

**ACPL ITEM
DISCARDED**

MCGRAW-HILL SERIES IN
Control Systems Engineering

*
621.8 G35c
Gibson & Tuteur 1058249

B & T

Control system components

PUBLIC LIBRARY

FORT WAYNE AND ALLEN COUNTY, IND.

STORAGE

DO NOT REMOVE
CARDS FROM POCKET

STO

**ACPL ITEM
DISCARDED**

2-25-59

CONTROL SYSTEM COMPONENTS

McGraw-Hill Series in Control Systems Engineering

JOHN R. RAGAZZINI AND WILLIAM E. VANNAH, *Consulting Editors*

COSGRIFF · Nonlinear Control Systems

GIBSON AND TUTEUR · Control System Components

GOODE AND MACHOL · System Engineering

LANING AND BATTIN · Random Processes in Automatic Control

RAGAZZINI AND FRANKLIN · Sampled-data Control Systems

SAVANT · Basic Feedback Control System Design

SMITH, O. J. M. · Feedback Control Systems

TRUXAL · Control Engineers' Handbook

Control System Components

JOHN E. GIBSON, Ph.D.

*Associate Professor of Electrical Engineering
Purdue University*

FRANZ B. TUTEUR, Ph.D.

*Associate Professor of Electrical Engineering
Yale University*

McGRAW-HILL BOOK COMPANY, INC.

New York

Toronto

London

1958

CONTROL SYSTEM COMPONENTS

Copyright © 1958 by the McGraw-Hill Book Company, Inc. Printed in the United States of America. All rights reserved. This book, or parts thereof, may not be reproduced in any form without permission of the publishers. *Library of Congress Catalog Card Number 57-12581*

THE MAPLE PRESS COMPANY, YORK, PA.

PREFACE

1C58249

Although a fairly large number of books dealing with the theory of servomechanisms has become available in recent years, most of them do not concern themselves to any great extent with the components required to implement the theory. The texts that have been written primarily about components have in general consisted merely of qualitative descriptions of available equipment. In the opinion of the authors of the present work a more quantitative approach to the subject of components is needed, and this book is an attempt to fill this need.

In making this attempt some compromises have, of course, been necessary. The number of components available for use in feedback control systems is so large that even a short description of each could easily fill several volumes. On the other hand it is possible to go into so much detail in the description and analysis of only a very small number of components that if the length of the book is to be kept within reason it becomes impossible to discuss many commonly used components at all. We have attempted to steer a middle course in this respect by presenting a fairly detailed but by no means exhaustive analysis of a number of what it is hoped are typical components. Then, in order to make the coverage of the field somewhat more complete, shorter descriptions of a somewhat larger group of devices have also been included. In most cases these are similar to the typical components discussed in detail. The decision on what was to receive detailed coverage, what was to receive only brief discussion, and what was to be left out altogether has, of course, been somewhat arbitrary. In general we have tried to exclude discussion of conventional devices that are used in standard fashion in control systems and we have considered in greatest detail those devices that seem to us to be of primary interest to the control-system engineer.

Our purpose has been to present engineering principles and methods of analysis rather than specific discussions of commercial devices. Catalogues of equipment and tables of data that may be found in standard handbooks have not been included. In a number of cases mathematical analyses are not completed; however, it is hoped that sufficient background has in all cases been developed so that the reader should have little difficulty in completing these analyses for himself. Usually, suggestions for such extensions will be found in the problems that follow each chapter.

The reader of this text is assumed to have a basic knowledge of servo theory or to be acquiring it concurrently. In particular, it is assumed that he has had an introduction into the Laplace transform method of analysis and is familiar with such concepts as transfer functions and complex impedance. These notions are reviewed briefly in Chapter 1 (mainly to introduce our notation which is slightly unconventional) and are assumed to be known in all of the other chapters. Also, a certain amount of electrical background is essential. For most of the topics covered in this book the electrical engineering background obtainable in one of the standard survey courses given to nonelectrical engineering majors should be sufficient. However, unfortunately for the non-electrical major, many of the components of control systems are electrical. Also, although the authors tried valiantly to keep the point of view of the nonelectrical control-system engineer in mind, they suffer from the handicap of being electrical engineers themselves. Hence it is clear that the greater the electrical background, the better. This applies particularly to Chapter 2 in which a knowledge of standard vacuum-tube theory is assumed.

On the whole the individual chapters of this book are self-contained and can be considered in any order, although naturally the authors prefer the order here followed. If the order is changed, it is advisable that the first twelve sections of Chapter 1, or at least Sections 1.10 to 1.12, be read first, since these sections contain the basic review material on the Laplace transformation and the Laplace transform nomenclature that is used throughout the book.

The authors have tried to treat each one of the major topics for the nonspecialist. For this reason a number of the chapters start with material that will appear elementary to the specialist and then proceed to more advanced levels. In general, the introductory sections containing the background material are kept as short as possible and no attempt at completeness has been made. Instead only the details necessary for the work at hand have been included. In most cases the central purpose of the analysis is the derivation of the transfer function of the component. For components described by nonlinear differential equations linearization techniques are presented and the usual assumptions that are made in order to obtain a linear approximation are discussed.

Large parts of certain chapters of the book have been presented in a graduate course on control-system components at Yale University for a number of years, while most of the other parts have also been presented in undergraduate and special courses offered to mechanical and electrical engineers in industry.

John E. Gibson
Franz B. Tuteur

CONTENTS

<i>Preface</i>	v
--------------------------	---

Chapter 1. Electric Networks for Control Systems	1
---	----------

1.1 Introduction	1.2 Resistors	1.3 Variable Resistors	1.4 The Precision Variable Resistor	1.5 Mechanical Characteristics of the Precision Variable Resistor	1.6 Nonlinear Precision Variable Resistors	1.7 Temperature Coefficient of Resistance	1.8 Capacitors	1.9 Inductors	1.10 The Analysis of Simple Networks	1.11 Steady-state Frequency Response from the Transfer Function	1.12 Asymptotic Representation of the Frequency Response	1.13 Common <i>RC</i> Networks	1.14 The Synthesis of <i>RC</i> Networks. Introduction	1.15 Properties of <i>RC</i> Networks	1.16 Properties of Two-terminal <i>RC</i> Networks	1.17 Transfer Functions Realizable with Four-terminal <i>RC</i> Networks	1.18 <i>L</i> Sections	1.19 Cascading of <i>L</i> Sections	1.20 Approximate Ladder-network Synthesis	1.21 Approximate Synthesis of <i>L</i> Sections	1.22 An Exact Synthesis Method of <i>L</i> Sections for Simple Transfer Functions	1.23 The Exact Synthesis of <i>L</i> Sections. General Method	1.24 Exact Synthesis of <i>RC</i> Driving-point Impedances	1.25 Numerical Example	1.26 Bridge-T and Twin-T Networks	1.27 The Use of Amplifiers with <i>RC</i> Networks	1.28 Reliability Problems
------------------	---------------	------------------------	-------------------------------------	---	--	---	----------------	---------------	--------------------------------------	---	--	--------------------------------	--	---------------------------------------	--	--	------------------------	-------------------------------------	---	---	---	---	--	------------------------	-----------------------------------	--	---------------------------

Chapter 2. D-C Amplifiers.	61
---	-----------

2.1 Introduction	2.2 Drift in D-C Amplifiers	2.3 Effect of Negative Feedback on Drift	2.4 Analysis of Simple Triode Circuits	2.5 The Cathode Follower	2.6 Input Impedance of the Cathode Follower	2.7 Drift Due to Component Variation	2.8 Cathode Follower with Plate-load Resistor	2.9 Equivalent Circuit for the Cathode Follower	2.10 The Triode Amplifier with Cathode Bias	2.11 The Difference Amplifier	2.12 The Miller Circuit	2.13 Phase Inverters	2.14 The Cathode-follower Inverter	2.15 The Paraphase, or Balanced, Inverter	2.16 The Cathode-coupled Inverter	2.17 Interstage Coupling Networks	2.18 The Transistor D-C Amplifier	2.19 Analysis of Simple Transistor Circuits	2.20 Single-stage Circuits	2.21 Drift in Transistor Amplifiers	2.22 Multistage Amplifiers	2.23 The Chopper-stabilized D-C Amplifier	2.24 Design Considerations	2.25 Triode Amplifier with Resistance-coupling Network	2.26 Design of Triode Amplifiers with Neon-tube Type Interstage Coupling Network	2.27 Design of Cathode-follower Circuits	2.28 Design of Difference Amplifiers	Problems
------------------	-----------------------------	--	--	--------------------------	---	--------------------------------------	---	---	---	-------------------------------	-------------------------	----------------------	------------------------------------	---	-----------------------------------	-----------------------------------	-----------------------------------	---	----------------------------	-------------------------------------	----------------------------	---	----------------------------	--	--	--	--------------------------------------	----------

Chapter 3. Power Amplifiers	115
--	------------

3.1 Introduction	3.2 A-C Power Amplifiers	3.3 D-C Power Amplifiers	3.4 Thyatron Amplifiers	3.5 Control of Firing Angle	3.6 The Thyatron
------------------	--------------------------	--------------------------	-------------------------	-----------------------------	------------------

Damper 8.6 Gears 8.7 Design of Gear Trains for Minimum Inertia 8.8
Gear Ratio for Load Matching 8.9 Backlash in Gears 8.10 Ball-screw
Actuator Problems

Chapter 9. Mechanical Components 334

9.1 Introduction 9.2 Differential Gears 9.3 The Universal Joint 9.4 Strain
Gauges 9.5 The Flyweight Tachometer 9.6 The Gyroscope. Introduction
9.7 The Gyroscope. Precession 9.8 Gyroscope. Equations of Motion 9.9
Practical Gyroscopes 9.10 Gyroscope Applications 9.11 Application of
Gyroscopes to Inertial Navigation Problems

Chapter 10. Pump-controlled Hydraulic Systems 363

10.1 Introduction 10.2 Basic Types of Hydraulic Control Systems 10.3 The
Pump-controlled Hydraulic System 10.4 Analysis of the Pump-controlled
System 10.5 Transfer Function of the Pump-controlled System 10.6 Other
Pump and Motor Types 10.7 Piston Pumps 10.8 Vane Pumps 10.9 Gear
Pumps 10.10 The Ball Pump 10.11 Hydraulic Transmission Lines Problems

Chapter 11. Valve-controlled Hydraulic Systems 386

11.1 Introduction 11.2 Spool-type Pilot Valves 11.3 Pulsed Operation of
Hydraulic Valves 11.4 Flow-pressure Relations in Orifices and Valves 11.5
Axial Hydraulic Reaction Forces in Spool-type Valves 11.6 Radial Hydraulic
Forces in Spool-type Valves 11.7 Graphical Analysis of Flow Control Valves
11.8 Linearized Small-signal Analysis 11.9 Flapper Valves 11.10 Nozzle
Valves 11.11 Slide Valves 11.12 Two-stage Valves 11.13 Pressure-regulat-
ing Devices 11.14 Choice of Operating Pressure for Hydraulic Systems Prob-
lems

Chapter 12. Pneumatic Systems 435

12.1 Introduction 12.2 Pneumatic Flow 12.3 Pneumatic Flow through an
Orifice 12.4 Compressible Flow in Pipes 12.5 Compressibility as an Equiva-
lent Spring 12.6 Pneumatic Transmission Lag 12.7 Pneumatic-Electric
Analog and Control Functions 12.8 A Pneumatic Control System Problems

Chapter 13. Pneumatic Components 461

13.1 Introduction 13.2 Comparison of Weight of Electric, Hydraulic, and
Pneumatic Systems 13.3 Choice of Operating Pressure 13.4 Pneumatic
Power Supplies 13.5 Pneumatic Control Valves 13.6 Pneumatic Motors
13.7 Actuators Problems

Index 481

CHAPTER 1

ELECTRIC NETWORKS FOR CONTROL SYSTEMS

1.1. Introduction. The vast majority of servo control systems in use at the present time employ electronic amplifiers and electric networks. This is true in spite of the facts that electronic components are often less reliable than other types and that conversion equipment is usually required to convert nonelectric signals into electric ones, and vice versa. The main reason for the popularity of electronic components is, of course, the relative ease with which electric signals can be manipulated. Thus, amplification of electric signals is easily accomplished by electronic circuits. While mechanical, hydraulic, or pneumatic amplifiers do exist and are widely used, they are much more complex and more expensive and, in general, perform much less well than electronic amplifiers. Similarly, mechanical, hydraulic, or pneumatic devices can be built to perform the function of frequency selection, i.e., to amplify some signals and to attenuate others on the basis of frequency. However, electric networks are ordinarily much preferred for this purpose, because they are easy to design and to construct and, except for certain specific instances,¹ their performance is superior to that of the other types of devices.

In the present chapter we deal with the electric networks used in control systems. These networks consist primarily of resistors and capacitors, although inductors are occasionally used. Before the networks are discussed, therefore, these three components are considered.

1.2. Resistors. Resistors are commonly made of a composition of carbon or from high-resistance wire. The fixed carbon resistor is manufactured in three standard power ratings: $\frac{1}{2}$ -, 1-, and 2-watt sizes; and in three standard tolerances: 20 (uncommon and no less expensive than 10 per cent tolerance), 10, and 5 per cent. Figure 1.1 gives the standard color code for resistance value and tolerance. It is becoming common practice to print the resistance on the body of the resistor in addition to color coding it.

The size of a carbon resistor is not indicative of its resistance value, since the composition is varied for various resistance values, but it does indicate the approximate power rating. Carbon resistors are made on a

¹ R. Adler, Compact Electromechanical Filter, *Electronics*, April, 1947, pp. 100–105. W. Roberts and L. Burns, Jr., Mechanical Filters for Radio Frequencies, *RCA Rev.*, vol. 10, pp. 348–365, 1949.

TABLE 1.1. STANDARD VALUES OF RMA CARBON RESISTORS

10 ohms	33 ohms
12	39
15	47
18	56
22	68
27	82
and powers of 10 thereof	

mass production basis, and the Radio Manufacturers' Association (RMA) has established standard production resistance values. These values are

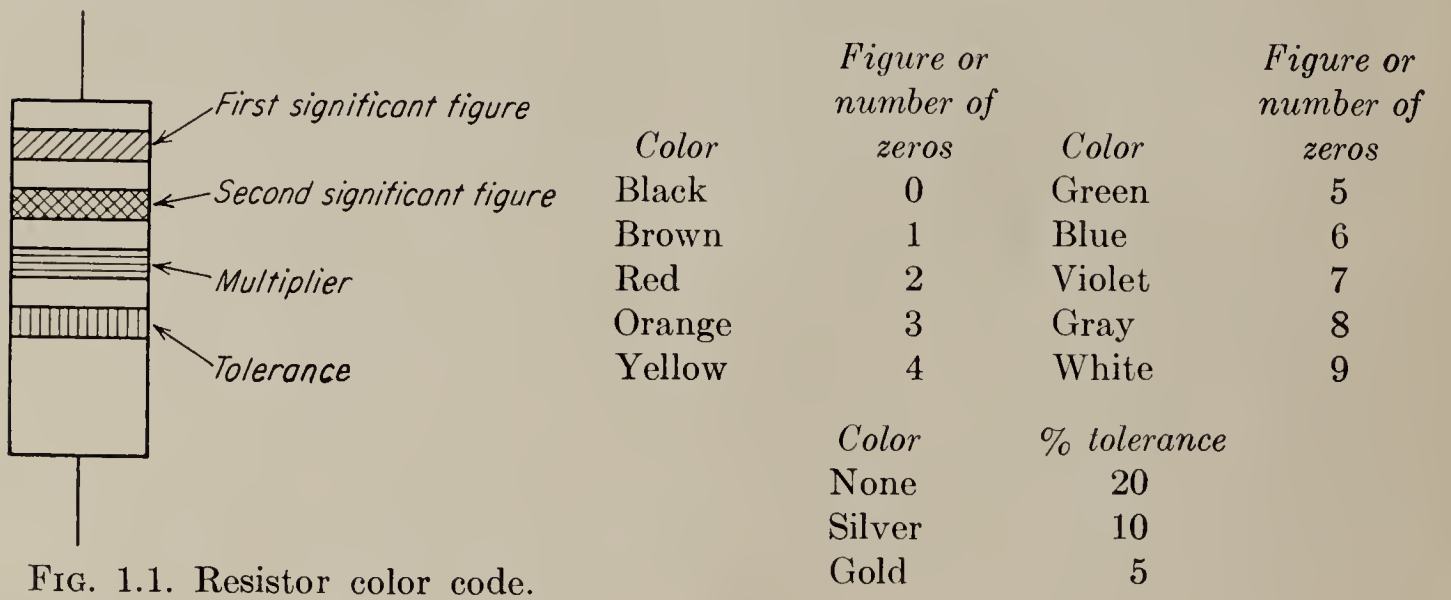


FIG. 1.1. Resistor color code.

given in Table 1.1. The values are chosen so that the percentage change in resistance between each of the values is approximately constant. Values at one-half the standard percentage step are also produced but are not so widely stocked and may be somewhat more expensive.

Wire-wound resistors may be obtained in any resistance value and wattage rating. Standard RMA values in the 5- and 10-watt ratings are commonly available from stock. Several companies specialize in supplying precision resistors to any accuracy required.

1.3. Variable Resistors. Sometimes the term “variable resistor” is applied to a device which has only two terminals. One terminal goes to one end of the resistor, while the other terminal is connected to the wiper arm. The symbol is given in Fig. 1.2a. The name “potentiometer” is then restricted to a device which has both ends of the resistance brought out to terminals and a third terminal for the connection to the wiper arm, as shown in Fig. 1.2b. This nomenclature is not universal, and both names may be used interchangeably.

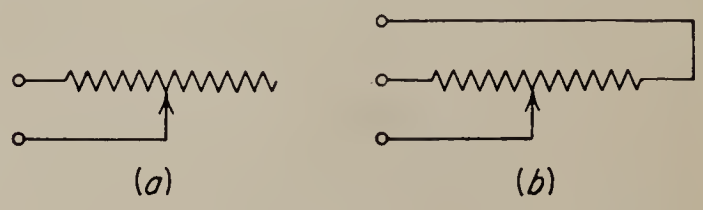


FIG. 1.2. Variable-resistor and potentiometer symbols.

Variable resistors either are made from a carbon composition or are wire-wound. The carbon resistor is less accurate and less expensive. The carbon variable resistor usually has a tolerance of 5 per cent of rated

value. The wire-wound variable resistor can be made to extremely accurate specifications. Since World War II a practically new industry has sprung up to manufacture variable resistors to standards which would have been unbelievable 10 years earlier.

1.4. The Precision Variable Resistor. The precision variable resistor, widely used in analogue computers and as a pickup device in servomechanisms, is manufactured in two forms: both are wire resistors. (A process which uses a deposited film of resistance material holds some promise, but it is not as yet of commercial importance.) The first form is the conventional wire-wound resistor held to close tolerances and with quality materials used. The resistance wire is wrapped on a “card,” or

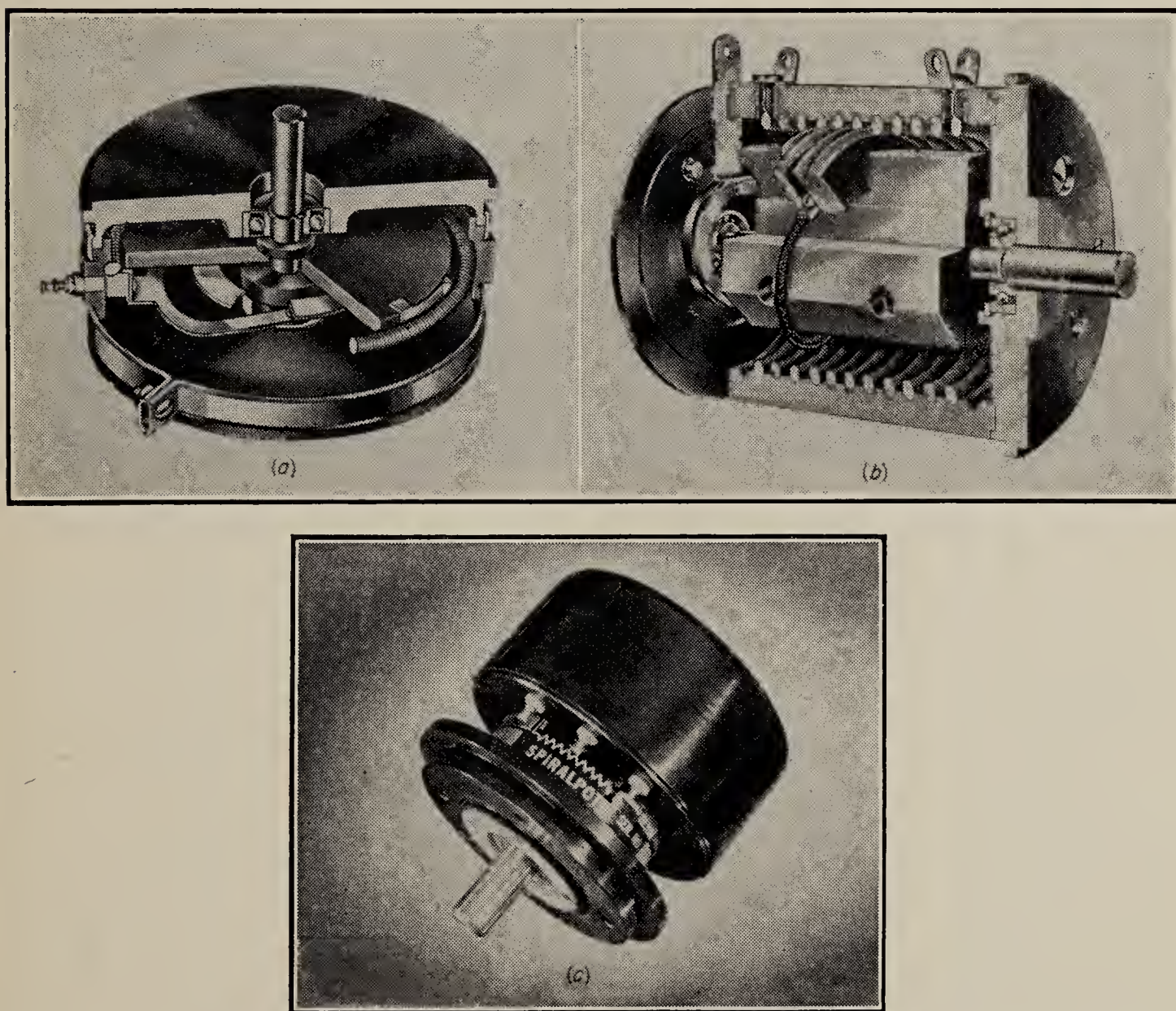


FIG. 1.3. Cutaway of precision variable resistors: (a) single-turn; (b) helical-wound resistor; (c) multiturn slide-wire. (Courtesy G. M. Giannini and Co.)

form, which is then bent into a circle. The wiper bears on the edge of the card and moves from turn to turn (see Fig. 1.3).

Accuracy of a variable resistor, or “linearity,” is commonly defined in two ways. So-called “best linearity,” “independent linearity,” or *normal linearity* is the maximum departure of the actual resistance curve from the best straight line which can be drawn through the actual curve, this departure being expressed as a percentage of the total resistance of the

device (see Fig. 1.4). Note that the straight line is not drawn through the origin, and thus a better figure for linearity is obtained than may be obtained if the zero-based linearity is used. *Zero-based linearity* is the maximum departure of the actual resistance curve from the best straight line drawn through the origin. This is the more natural definition, but precision-resistor manufacturers have adopted normal linearity almost universally. A single-turn wire-wound resistor may be obtained from stock with 0.1 per cent normal linearity for resistance values of 5,000 ohms and above. The body diameter of this device is of the order of 3 in. If either the resistance or the body diameter is smaller, the number of turns of wire must be reduced, and thus the linearity is reduced.

The resolution of such a device is also important. *Resolution* of a variable resistor is defined as the smallest change in resistance which can be detected, expressed as a percentage of the total resistance of the device.

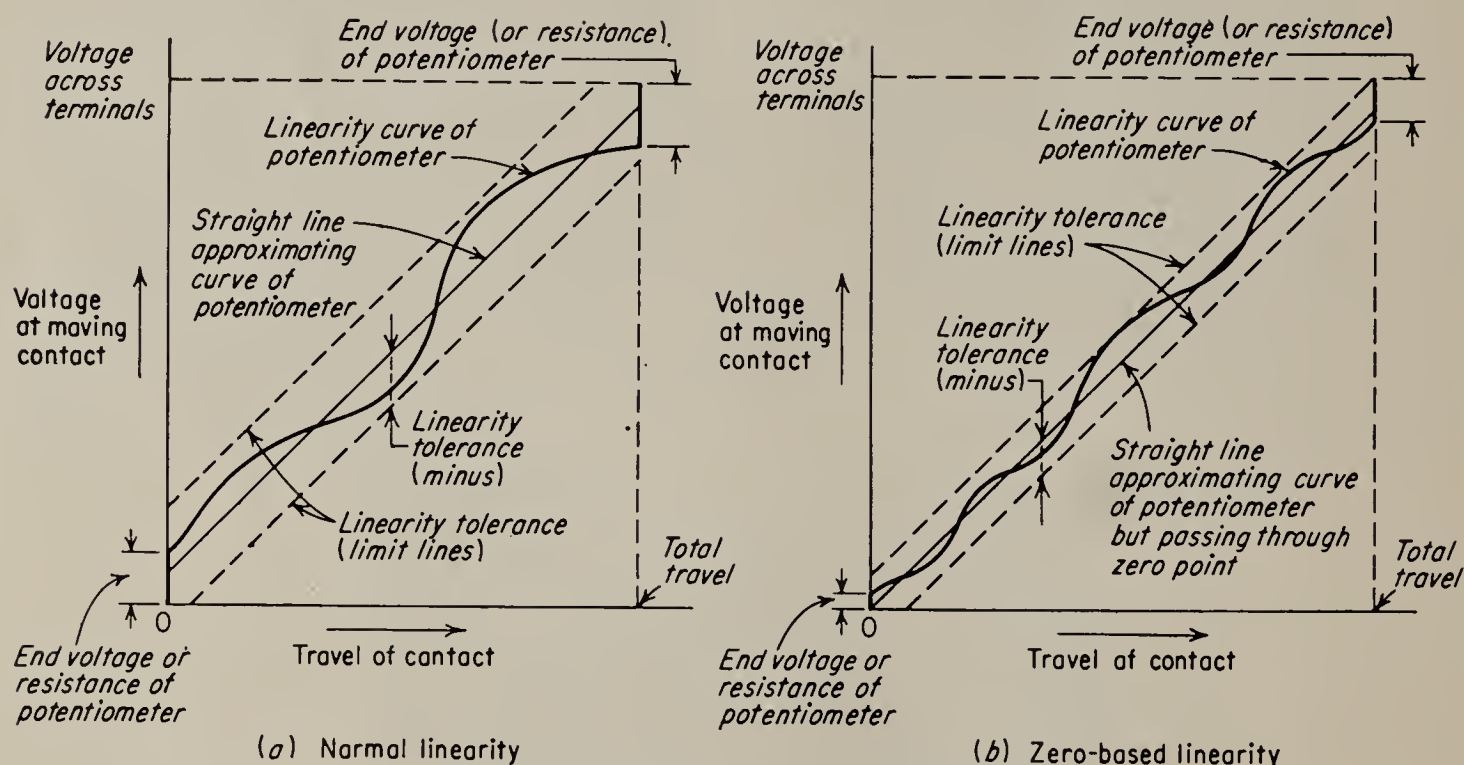


FIG. 1.4. Linearity of variable resistors.

The smallest resistance change detectable in a wire-wound resistor is that of one turn of wire, and thus

$$\text{Resolution} = \frac{1}{\text{total number of turns of wire}} \times 100\% \quad (1.1)$$

Typically, a single-turn (360° card) wire-wound resistor with a body diameter of 3.5 in. has a resolution of about 0.04 per cent at 10,000 ohms. A larger body diameter or higher resistance would result in better resolution (a smaller percentage).

It is possible to construct a wire-wound resistor with the card not bent into a circle but rather with the wire spiraled in the form of a helix. This allows more turns of resistance wire and increases the linearity and improves the resolution. While the usual number of turns of the helix is

10, resistors are commercially available with as many as 40 or 50 helical turns. Figure 1.3*b* shows a 10-turn helical resistor.

In another type of variable resistor the resistance is not wound on a card so that the wiper moves from turn to turn of resistance wire; instead the wiper is always in contact with the resistance wire. The high-resistance wire is shaped in a spiral, usually of 10 turns. This is merely a modification of the linear slide-wire potentiometer. The advantages of the slide-wire resistor are infinite resolution (0 per cent by the standard definition), low capacitive and inductive effects, and excellent zero-based linearity. The disadvantages are the difficulty of maintaining linearity for resistance values above 50,000 ohms and the somewhat higher cost. Slide-wire variable resistors are available in resistance values up to 25,000 ohms with zero-based linearity as good as 0.025 per cent and normal linearity as good as 0.01 per cent. The zero-based linearity of the slide-wire device excels that of the wire-wound resistor in general, and the normal linearity of the slide-wire resistor compares favorably with that of the wire-wound resistor in the resistance range of 10,000 to 25,000 ohms. At lower values of resistance the normal linearity of the slide-wire resistor is better than that of the wire-wound device.

1.5. Mechanical Characteristics of the Precision Variable Resistor. Precision resistors are usually mounted within a steel, aluminum, or plastic case. The shaft bearings on the standard models are precision ball bearings which require no lubrication. Most models will withstand accelerations of up to 100 times that of gravity along any axis without damage. The multiturn models are equipped with mechanical stops at each end of the resistance wire. Single-turn units may be equipped with mechanical stops, or in some models continuous mechanical rotation is possible. Standard models with ball bearings require about 1 oz-in. starting torque, and low-torque models with ball bearings are available which require less than 0.25 oz-in. starting torque. Models are commercially available with jewel bearing which require a maximum of 0.003 oz-in. starting torque. Usually the starting torque may be divided by 2 to find the running torque required.

Many manufacturers do not seem to realize that, in addition to the quantities mentioned above, the control engineer requires the moment of inertia of the rotating parts of the variable resistor, in order to compute the mechanical time constant of the unit into which the resistor is connected. Therefore this datum is not always available. Assuming the usual lightweight construction, a resistor with a body diameter of 3 in. may have a moment of inertia of about 0.05 oz-in.². A resistor with a body diameter of less than 1 in. may have a shaft moment of inertia of less than 0.005 oz-in.², and one manufacturer has a regular production model with a starting torque of 0.005 oz-in. and a moment of inertia of 5×10^{-6} oz-in.². Many standard models have fittings which make

it possible to connect several resistors coaxially on the same shaft; this is called "ganging" the resistors. If three or more resistors are ganged, the shaft should be brought out through the last resistor and supported by a bearing to prevent twisting of the first resistor's shaft by the weight of the ganged resistors.

1.6. Nonlinear Precision Variable Resistors. It is sometimes necessary to cause the voltage at the resistor wiper arm to be a particular nonlinear function of shaft rotation. This might be necessary, for instance, in synthesizing the optimum switching function in a relay servo or in the scheduling control for a jet fuel-control system. One method of accomplishing this is to wire the resistance card deliberately to give the required function. The disadvantages of this method are that it is very difficult to construct the winding accurately and, second, that the resolution varies with the spacing of the resistance wire unless the width of the card is varied in accordance with the function. A second method, which is more practical, consists in arranging a linear resistor with taps every few degrees. It is impossible to locate taps any closer than 10 mechanical degrees to each other, and these taps must be installed at the factory during the manufacturing process. Padding resistors may be installed across the taps to shape the resistance characteristic in a series of straight-line segments that approximate the desired curve. This is called *resistance padding*. The linear-resistor characteristic must have at least as steep a slope as the steepest slope of the desired characteristic, since it is impossible to increase the slope of the linear resistor by placing a padding resistor in parallel. Note further that the slope can never be less than zero, since this would require negative resistance.

A second method of padding the linear resistor to approximate a curve with straight-line segments is called voltage padding. *Voltage padding*¹ consists in placing voltages across the taps of the linear resistor to shape the characteristic of the resistor's output voltage versus shaft position to approximate a desired nonlinear function; for typical example see Fig. 1.5. Voltage padding is more flexible than resistance padding because there are no restrictions on the slope of the desired characteristic. Theoretically, it is merely necessary to calculate the voltage required between taps and to connect a suitable voltage source between the taps. Actually, several other factors must be considered. First, it is necessary that the power-dissipation rating of the variable resistor not be exceeded. Not only must the over-all rating be observed, but also the power dissipated between any two taps must not be excessive. The minimum total value of resistance of the variable resistor is established by the segment with the maximum voltage gradient.

¹ G. R. Korn, Design and Construction of Universal Function Generating Potentiometers, *Rev. Sci. Instr.*, vol. 21, no. 1, p. 77, January, 1950.

A second factor that must be considered is the internal impedance of voltage supplies that are placed across the taps. The simplest method of obtaining the proper voltage gradient is to use a fixed battery and a voltage-dropping resistor. A more sophisticated approach is to employ a regulated, adjustable voltage supply that has a very low output impedance.

Finally, the current drawn by the wiper arm must be considered. Frequently the potentiometer resistance is made low enough that the voltage drop produced by the wiper current is negligible. Alternatively, if the resistance connected between the wiper arm and ground is constant, it may be more accurate to take the loading into account in the design and to adjust the padding in such a way that the loaded potentiometer has the desired characteristic.

It is possible to combine the several methods of obtaining nonlinear resistors. For instance, voltage padding may be used between a few taps and then the characteristic between the voltage pads may be shaped by resistance padding. Resistance padding can only decrease the volts per degree of the characteristic (i.e., make it closer to zero), and therefore the unpadded resistor must be assigned a slope at least as steep as the greatest slope of the desired characteristic, and voltage padding must be used to establish negative slopes.

Most manufacturers of precision variable resistors are willing to undertake the design and construction of nonlinear variable resistors to the customer's specifications.

1.7. Temperature Coefficient of Resistance. Most physical materials exhibit some variation of resistance with temperature. On the sub-microscopic level, resistance is presumably the interference with the flow of electrons, or current, by the random motion of atomic particles of the material. At absolute zero temperature there is no random motion of these particles, and thus we would expect the material to have zero resistance there. This is borne out, in general, by experimentation, but as in most simplifications there are exceptions. At very low temperatures the curve of resistance does not approach zero in a straight line, and, most particularly, there are certain materials, such as carbon, whose resistance increases with a decrease in temperature in the normal operating range.

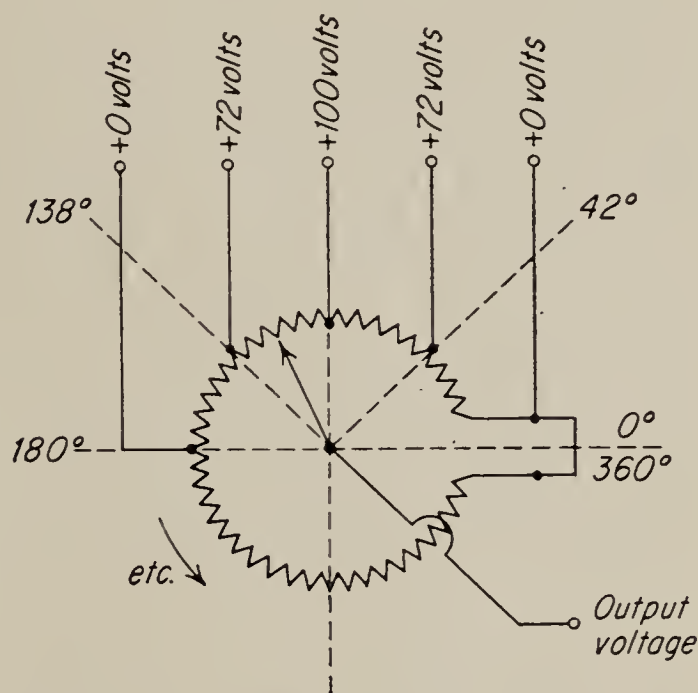


FIG. 1.5. Voltage padding designed to generate a sine function.

The empirical relation for the temperature variation is discussed in most elementary texts on electricity¹ and is given in Eq. (1.2),

$$R_h = R_c[1 + \alpha(t_h - t_c)] \quad (1.2)$$

where R_h = hot resistance

R_c = cold resistance

t_h = hot temperature

t_c = cold temperature

α = temperature coefficient of resistance

Tables for α for various materials are given in engineering handbooks. As a result of the way in which α is defined, it is dependent not only on the material, but also on the cold temperature at which it is computed. Usually the tables list the α computed for a cold temperature of 20°C. This value is usually adequate for most calculations. However, should a very exact calculation be necessary for a cold temperature other than 20°C, a new α may be computed by setting up a linear ratio.

The temperature coefficient of resistance of the material used in making precision variable resistors is a very important consideration. In the lower resistance values a low-nickel-high-copper alloy is usually used. It has a resistivity of 60 ohms/unit volume and a temperature coefficient of 0.0071/°C. In the intermediate resistance values, Advance Alloy (50-50 copper nickel), with a temperature coefficient of 0.00002 and a resistivity of 294, is used. In the higher resistance values, Nichrome V (nickel chrome), with an α of 0.00013 and a resistivity of 650, is normally used. Karma wire, with a resistivity of 800 and an α of 0.00002, is sometimes used in high-resistance elements, although its life and corrosion characteristics are not quite so good as those of Nichrome V.²

1.8. Capacitors. Several types of capacitors are of commercial importance. The *air-dielectric variable* capacitor has one set of plates mounted on a shaft which may be turned with respect to the other set of plates. The shaft is eccentric with the fixed set of plates, and as it is turned, the effective spacing between the two sets of plates is varied, thus varying the capacitance. The air-dielectric variable capacitor is more commonly used in the tuning circuits of radios and electronic devices than in control systems.

The *tubular* capacitor, or “paper” capacitor, is widely used in standard electronic circuits. It consists of two thin conducting sheets of tin foil or silver foil separated by a sheet of dielectric. Usually the dielectric is impregnated paper. The sheets are then rolled into a compact tube, using one additional sheet of dielectric to prevent shorting the conducting

¹ See, for example, Cook and Carr, “Elements of Electrical Engineering,” John Wiley & Sons, Inc., New York, 1951.

² Data furnished by the Helipot Corporation, South Pasadena, Calif.

plates. The tubular capacitor is an inexpensive element, but it should not be used in critical circuits. The dielectric tends to deteriorate with age, and the capacitor is not stable with respect to temperature and humidity. The tubular capacitor may exhibit an appreciable leakage current in the larger sizes. Leakage current is the current that flows from one plate to the other in a capacitor if the dielectric is not perfect. The leakage path may be represented approximately as a resistor in parallel with the capacitor.

Recently a new silicon dielectric material has been used in tubular capacitors. This new dielectric results in a physically smaller capacitor and reportedly improves to some extent the undesirable characteristics mentioned above.

The *oil-filled* capacitor provides more capacity in the same physical space than does the tubular capacitor but, because of the metal case required, it is more expensive. For capacities of greater than $0.5\ \mu\text{f}$ and for rated voltages of greater than 100 volts the oil-filled capacitor is generally used, since it is difficult to make paper or mica capacitors in these large sizes.

The *mica* capacitor is more dependable and more expensive than the tubular capacitor and is available in the same ratings. The construction consists of alternate layers of metal foil and mica as a dielectric. Every other metal plate is connected together, and the assembly is incased in bakelite.

High capacities in a small space are available in the *electrolytic* capacitor. However, the usual electrolytic capacitor is polarized, i.e., only voltages of one polarity may be applied to it without damage. The electrolyte is a solution of boric acid which acts as the negative plate. An aluminum anode acts as the positive plate, while a thin film of aluminum oxide which forms on the anode acts as the dielectric. In the so-called "dry" electrolytic capacitor, the electrolyte is damp paper soaked in boric acid or ethylene glycol. The solution will ionize and break down if a reverse-polarity voltage is applied to it. This makes the usual electrolytic capacitor useless for a-c-voltage applications. The polarized electrolytic capacitor is used in d-c power-supply filters and in bypass applications in electronic amplifiers.

It is possible to construct an electrolytic capacitor with two anode plates immersed in a common electrolyte, for a-c service. This non-polarized capacitor is equivalent to two polarized electrolytic capacitors connected in series, i.e., back to back, with the negative terminals common. The anode area must be four times the area of a polarized capacitor of the same capacity. The nonpolarized electrolytic capacitor is useful in applications where the exact value of capacitance is not important, as in phase-splitting service for single-phase, a-c induction motors.

1.9. Inductors. Unlike resistors and capacitors, inductors are not commonly mass produced to standard specifications. This is due to the extremely wide range of parameters and ratings that are encountered.

Air-cored inductors are usually made only in sizes up to a few millihenrys, since they become very bulky for larger inductance values. For this reason air-cored inductors are little used except in high-frequency communication applications.

Iron-cored inductors are smaller and more compact than air-cored devices and are available in inductance values of up to several hundred henrys. In the past decade very-high-permeability alloys such as Deltamax, permalloy, and Supermalloy have become available for use as core material. Deltamax typically will support 90,000 maxwells/in.² with about 0.4 amp-turn/in. at the knee of the magnetization curve. Permalloy saturates at 39,000 maxwells/in.² with about 0.2 amp-turn/in., and Supermalloy saturates at 42,000 maxwells/in.² with 0.12 amp-turn/in. From the intense interest in improved core materials for magnetic amplifiers further advances are to be expected in magnetic steels. Inductors using these newer, sharply saturating materials must be designed to operate below the knee of the curve. Grain-oriented core materials conventionally are sold in U-shaped laminations or in the form of rolls of "tape." In either case, bending, squeezing, or heating will damage the magnetic properties. Powdered-iron toroids are also used extensively and result in inductors having extremely low iron losses.

Although practical resistors and capacitors approach mathematical perfection, a "pure" inductance cannot be constructed. Inductors are subject to copper loss from the resistance of the wire with which they are wound. Furthermore, the iron core results in iron losses due to hysteresis and eddy currents. A very important parameter that must be considered in choosing an inductor for a given application is, therefore, the quality factor Q , defined as the ratio of power handled to power dissipated. It may be shown¹ that for inductances this is approximately equivalent to

$$Q = \frac{\omega L}{R} \quad (1.3)$$

where ω = frequency, radians/sec

L = inductance, henrys

R = equivalent resistance including the iron core loss, ohms

This definition of Q indicates properly that the performance of inductors is poor at low frequencies, but it does not show that performance is also poor at high frequencies. As a result of interwiring capacitance, coils resonate at some frequency, and the performance deteriorates

¹ Ramo and Whinnery, "Fields and Waves in Modern Radio," John Wiley & Sons, Inc., New York, 1953.

rapidly at frequencies above this value. Modern toroidal coils carefully made on high-quality cores of about 1 in. diameter will exhibit a peak Q of about 200 in a frequency range of 2 to 5 kc. The Q falls rapidly above and below the peak frequency. Typically the Q would be 100 at 10 kc and 1 kc and would be less than 5 at 100 cps.

Since it is impossible to obtain a high- Q coil for operation below 100 cps, coils are seldom deliberately used as circuit elements in d-c servos. Of course, this statement does not apply to the inherent inductance of motor windings and solenoids, etc. In a-c servos, where a 400-cps carrier is often used, inductances are occasionally employed. Even here, however, the size, weight, difficulty in shielding, low quality factor, expense, non-reproducibility in mass production, and nonlinearities argue against the use of inductors.

1.10. The Analysis of Simple Networks. Networks are analyzed by the application of Kirchhoff's voltage and current laws. Since most networks used in control systems have a common ground connection and seldom contain mutual-inductance elements, the node method of analysis¹ normally results in simpler equations and yields the results with less labor than the mesh method. The result usually of primary interest in control-system applications is the network *transfer function*, i.e., the relation between input and output voltage.

Although it is not our purpose in this text to present an extensive discussion of network analysis or of Laplace transform techniques, a simple example will be worked out, partly by way of review, partly to establish the system of nomenclature used in later chapters.² The network considered is the simple RC "rate network" shown in Fig. 1.6.

To analyze this network by means of the node method, we equate the currents leaving the node e_o to zero. It is frequently assumed that any device connected to the output terminals of the network does not "load" the network; in other words, no current flows to the right from the node e_o . With this assumption we have,

$$C \left(\frac{de_o}{dt} - \frac{de_i}{dt} \right) + \frac{e_o}{R} = 0 \quad (1.4)$$

¹ Gardner and Barnes, "Transients in Linear Systems," John Wiley & Sons, Inc., New York, 1942, pp. 38-43.

² Readers not familiar with the Laplace transform method for analyzing networks are referred to Gardner and Barnes, *ibid.*, and also to Chestnut and Mayer, "Servomechanisms and Regulating System Design," John Wiley & Sons, Inc., New York, 1951, vol. 1, pp. 66-96. For a mathematical treatment of the Laplace transformation, see Doetsch, "Theorie und Anwendung der Laplace-transformation," Dover Publications, New York, 1943.

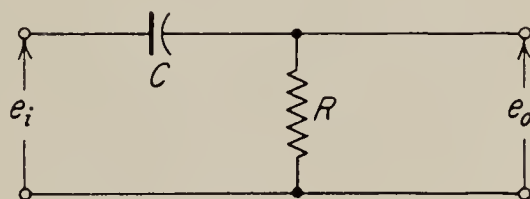


FIG. 1.6. RC rate network.

The transfer function of the network is defined as the ratio of the Laplace transform of the output voltage to that of the input voltage with all initial conditions set to zero. Accordingly, we transform Eq. (1.4) to obtain

$$Cs[E_o(s) - E_i(s)] + \frac{1}{R} E_o(s) = 0 \quad (1.5)$$

We use the shorthand notation $\hat{e} \triangleq E(s)$; hence Eq. (1.5) may be rewritten in the form

$$Cs(\hat{e}_o - \hat{e}_i) + \frac{1}{R} \hat{e}_o = 0 \quad (1.6)$$

Solving for \hat{e}_o , we obtain the transfer function

$$\frac{\hat{e}_o}{\hat{e}_i} = \frac{RCs}{RCs + 1} \quad (1.7)$$

It may be shown¹ that the quantity RC has the dimensions of time. It is, therefore, referred to as a *time constant* and given the symbol T . With this notation, Eq. (1.7) can be written in the equivalent form

$$\frac{\hat{e}_o}{\hat{e}_i} = \frac{Ts}{Ts + 1} \quad (1.8)$$

Equations (1.7) and (1.8) are in the *time-constant form*. This form is identified by the fact that the constant term is unity. The time-constant form is one of two standard forms in which transfer functions are commonly expressed. The other common form, which may be referred to as the *root form*, is given for the simple network discussed here by

$$\frac{\hat{e}_o}{\hat{e}_i} = \frac{s}{s + (1/RC)} = \frac{s}{s + (1/T)} \quad (1.9)$$

This form is usually more convenient when the Laplace transform is to be inverted to obtain the time function to which it corresponds.

The transfer function for simple networks can usually be obtained more rapidly by using the principle of the voltage divider. This principle states that in a series circuit the ratio of the voltage across any element or group of elements to the total applied voltage is equal to the ratio of the

¹ C is defined as charge/volt, and R is defined as volts/amp; thus, dimensionally,

$$(R)(C) = \frac{\text{volts}}{\text{amp}} \times \frac{\text{charge}}{\text{volt}} = \frac{\text{charge}}{\text{amp}}$$

and current is the rate of flow of charge, or charge/time; thus

$$RC = \frac{\text{charge}}{\text{charge/time(sec)}} = \text{time (sec)}$$

impedance of the element or elements to the total impedance of the circuit. In order to apply this principle, use is made of the concept of generalized impedance, defined as the ratio of the Laplace transform of the voltage across an element to the Laplace transform of the current in the element. Thus, the impedance of a capacitor is $1/Cs$, that of an inductance is Ls , and that of a resistance is R . Using these ideas, we have for our network

$$\frac{\hat{e}_o}{\hat{e}_i} = \frac{R}{R + (1/Cs)} = \frac{RCs}{RCs + 1} \quad (1.10)$$

The transfer function is thus obtained in a single step.

1.11. Steady-state Frequency Response from the Transfer Function. The Laplace transform method of analysis is perfectly general. There

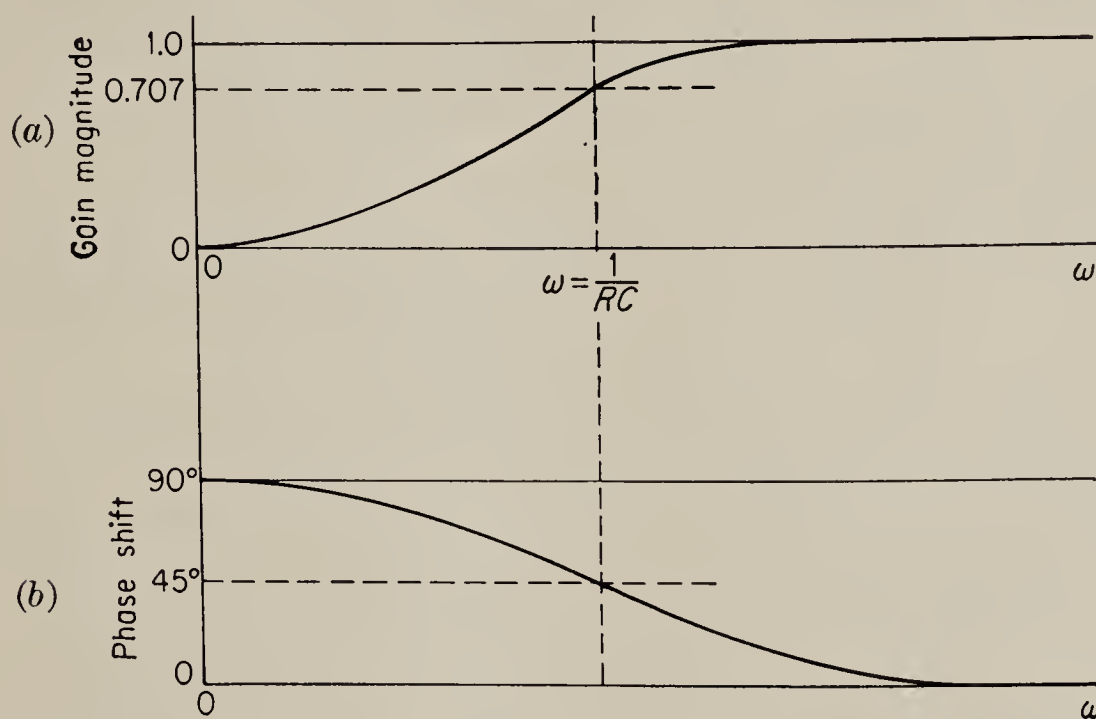


FIG. 1.7. (a) Magnitude and (b) phase angle of the transfer function for the circuit shown in Fig. 1.6.

have been no limitations placed on the voltage that is applied to the circuit except that it exists only for positive time. If we are willing, however, to restrict ourselves to sinusoidal voltages and to steady-state operation, we may solve for the response of any circuit to any frequency given merely the transfer function of the circuit. It is only necessary to substitute $j\omega$ for s ($j = \sqrt{-1}$) and then to let ω be any value of frequency in which we are interested.¹ The magnitude of the resulting complex number is the amplitude ratio of output to input, and the angle is the phase difference between output and input. For the circuit considered in Fig. 1.6 this information is available in Eq. (1.11). The complex numbers in the numerator and denominator are converted to polar form and divided. The resultant magnitude is shown in Fig. 1.7a, while the angle between input and output is shown in Fig. 1.7b.

¹ See Chestnut and Mayer, *op. cit.*, pp. 99–103.

$$\frac{\bar{e}_o^*}{\bar{e}_i} = \frac{j\omega T}{1 + j\omega T} = \frac{\omega T / 90^\circ}{\sqrt{1 + (\omega T)^2} / \tan^{-1} \omega T} = \frac{\omega T}{\sqrt{1 + (\omega T)^2}} \angle 90^\circ - \tan^{-1} \omega T \quad (1.11)$$

1.12. Asymptotic Representation of the Frequency Response. The transfer functions of networks with lumped parameters such as R 's, L 's, or C 's always consist of the ratio of products of simple factors such as s or $Ts + 1$. In the simple example of Fig. 1.6 the transfer function consists only of the simple ratio of two factors of this type, but, as will be shown presently (Sec. 1.16), more complicated networks will in general contain more of these factors in both the numerator and the denominator. For this reason it is convenient to represent the amplitude characteristic given in Fig. 1.7a on a log-log scale. This enables one to add or subtract the characteristic curves of the typical factors to build up the total characteristic. Logarithmic representation also permits showing a wide range of variables in a small space.

The logarithmic representation of a factor of the type $Ts + 1$ is simplified by first considering its asymptotic behavior for very small and very large $s = j\omega$.

For $\omega T \ll 1$, we have

$$\log |1 + j\omega T| \approx \log 1 = 0 \quad (1.12)$$

while, for $\omega T \gg 1$,

$$\begin{aligned} \log |1 + j\omega T| &\approx \log \omega T \\ &= \log \omega + \log T \end{aligned} \quad (1.13)$$

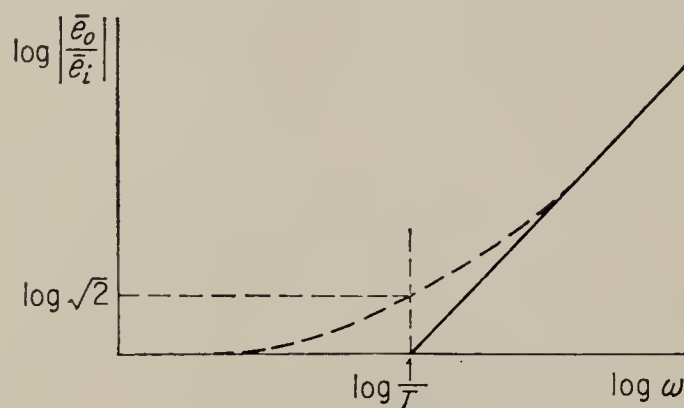


FIG. 1.8. Asymptotic representation of $\log |1 + j\omega T|$.

Thus the asymptote for low frequencies is the log- ω axis, while for high frequencies the asymptote is a straight line with a positive slope of 1, intersecting the log- ω axis at the point $\omega = 1/T$. These two asymptotes are shown in Fig. 1.8 as the solid lines. The actual characteristic $\log |1 + j\omega T|$ is shown dotted. Note that the actual curve does not depart very far from the asymptotes at any point. For real T the maximum departure occurs at the point where the two asymptotes meet. This is the so-called “break point,” for which $\omega = 1/T$. At this point the value of the actual characteristic is $\sqrt{2}$; hence the departure of the asymptotes from the true curve is of the order of 30 per cent. Since the asymptotes represent the actual curve so closely, it is common practice to draw only the asymptotes to represent the amplitude characteristic of a transfer function. The resulting curve is the asymptotic amplitude diagram.

* The bar indicates that $j\omega$ has been substituted for s .

For the simple network of Fig. 1.6 we obtain the asymptotic representation by taking the logarithm of the magnitude of the frequency ratio [Eq. (1.8) in which $s = j\omega$]. We obtain

$$\log \left| \frac{\bar{e}_o}{\bar{e}_i} \right| = \log T + \log \omega - \log |1 + j\omega T| \quad (1.14)$$

Thus there are three components. The asymptotic characteristics for the three components are shown in Fig. 1.9a, while the total characteristic for the rate network of Fig. 1.6 is shown in Fig. 1.9b. The dotted line again represents the actual curve.

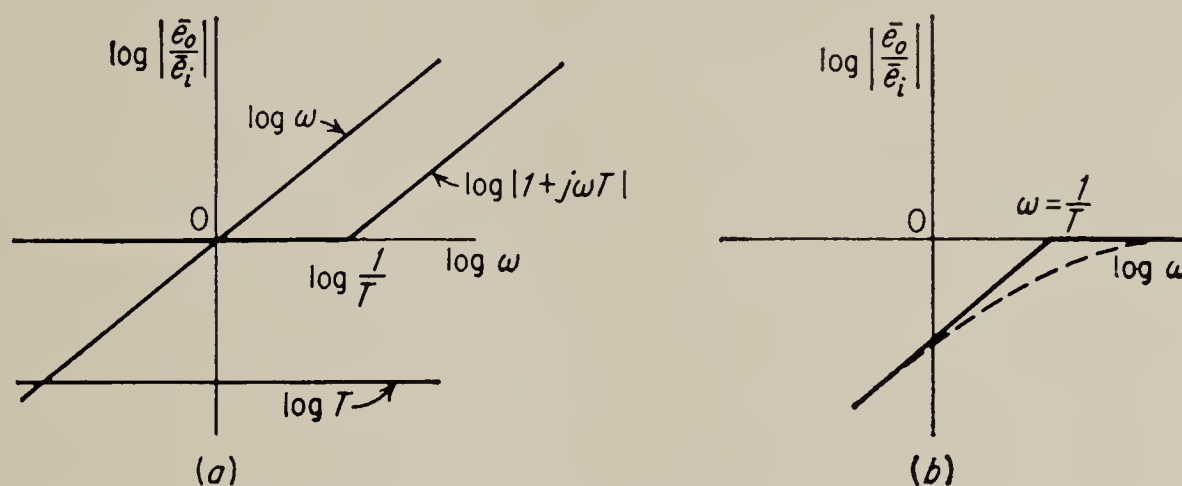


FIG. 1.9. Asymptotic representation of rate-network transfer function.

For more complicated functions the procedure is exactly the same, except that there are more factors whose characteristics must be combined. Under these conditions a somewhat simpler technique for drawing the asymptotic diagram is convenient. It is based on the following rules:

1. Start with a frequency low enough that, for all the factors of the form $1 + Ts$, $|Ts| < 1$. As far as the asymptotic representation is concerned, all of these factors then are equal to unity.

2. The low-frequency asymptote may now be drawn.

3. As the frequency is increased, the T 's product in some factors will become greater than unity. If such a factor appears in the numerator, the slope of the asymptotic diagram increases by 1, while if it is in the denominator, the slope decreases by 1. If there is more than one factor having the same T , the slope is increased (or decreased) by the number of such factors appearing.

4. The time constants T need not always be real numbers. If they are complex, they must occur in complex-conjugate pairs. As far as the asymptotic diagram is concerned, a pair of factors with complex-conjugate time constants is treated exactly as if they were a pair of factors with equal real time constants; i.e., they result in a slope change of ± 2 . However, since complex time constants indicate a resonance condition in the transfer function, the actual curve may depart very markedly from

the asymptotic curve near the break point. The reader is referred to texts on servomechanism theory for a complete discussion of this problem.¹

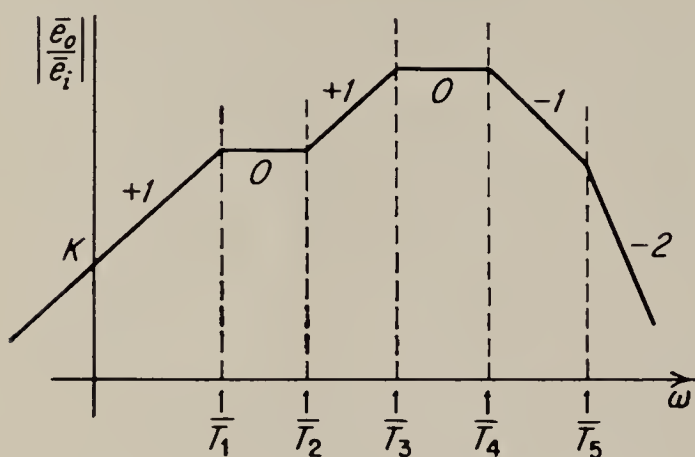
The procedure is best illustrated by an example. Thus, consider the transfer function

$$\frac{Ks(T_2s + 1)}{(T_1s + 1)(T_3s + 1)(T_4s + 1)(T_5s + 1)} \quad 1 > T_1 > T_2 > T_3 > T_4 > T_5 \quad (1.15)$$

For $s = j\omega < 1/T_1$, the asymptotic representation is identical to that of the transfer function Ks . It has a $+1$ slope, and for $s = 1$ its level is K . As the frequency increases to $\omega = 1/T_1$, the first denominator term becomes effective, and the slope changes to zero. When $\omega = 1/T_2$, the

slope changes to $+1$, etc. The complete curve is shown in Fig. 1.10.²

The magnitude of the transfer function is sometimes given in decibels (db). When used in connection with transfer functions, the decibel value is usually defined as



$$\text{db} = 20 \log \left| \frac{e_{\text{out}}}{e_{\text{in}}} \right| \quad (1.16)$$

FIG. 1.10. Typical asymptotic diagram.

The decibel expression is no more convenient in use than any other

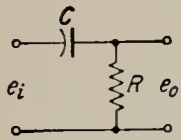
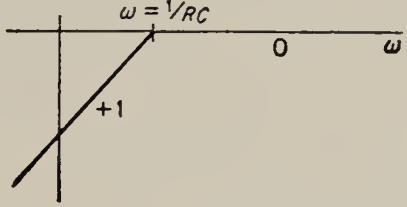
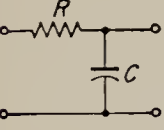
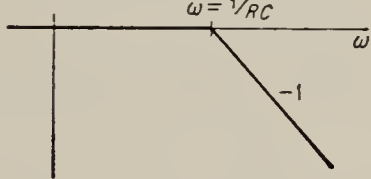
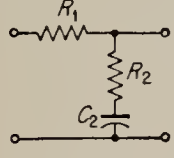
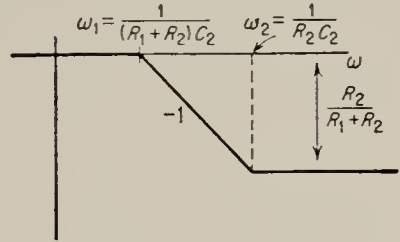
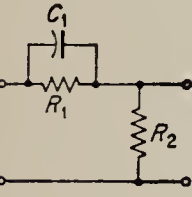
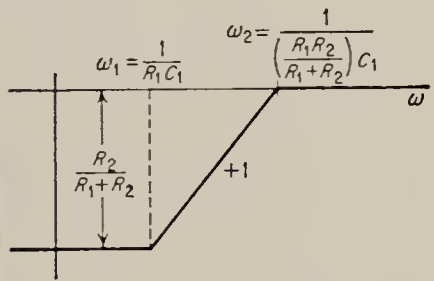
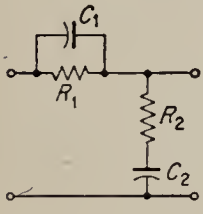
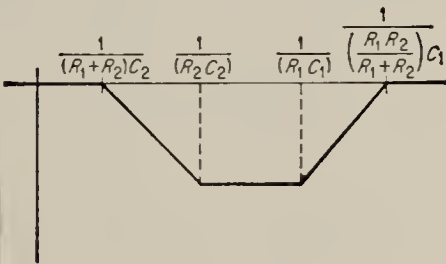
logarithmic ratio. Since the product of two numbers is equivalent to the sum of their logarithms, addition is substituted for multiplication. The use of the actual numerical ratios combined with the log-log plot will be found more convenient than the decibel function in the design of corrective networks. The use of the actual ratio makes it possible to read required information directly from the plot without conversion, and furthermore the ambiguity in the definition of the decibel is avoided.

1.13. Common RC Networks. The five simple *RC* networks given in Table 1.2 are very commonly used in servo systems, and for many of the simpler systems no other networks are required. The table was constructed by first drawing the physical network configuration and then analyzing it, as shown in Sec. 1.10 above. In all cases the assumptions are made that the driving source has zero output impedance and that the

¹ James, Nichols, and Phillips, "Theory of Servomechanisms," Radiation Laboratory Series, vol. 25, McGraw-Hill Book Company, Inc., New York, 1947, pp. 144, 176. Also Chestnut and Mayer, *op. cit.*, pp. 302-315.

² The logarithmic plot may be made by plotting the logarithms of the values on linear graph paper, or by plotting directly the values scaled logarithmically. The latter procedure is universal; thus, in the remainder of the text we shall label the axes with the functions rather than their logarithms.

TABLE 1.2. TABLE OF COMMONLY USED RC CORRECTIVE NETWORKS

Physical configuration	Names	Transfer function	Asymptotic amplitude characteristic vs. frequency
	Rate network; high-pass filter	$\frac{e_o}{e_i} = \frac{RCs}{RCs + 1}$	
	Reset network; integral controller; low-pass filter	$\frac{e_o}{e_i} = \frac{1}{RCs + 1}$	
	Proportional plus reset network; lag network	$\frac{e_o}{e_i} = \frac{R_2 C_2 s + 1}{(R_1 + R_2) C_2 s + 1}$	
	Proportional plus rate network; lead network	$\frac{e_o}{e_i} = \frac{R_2}{R_1 + R_2} \frac{R_1 C_1 s + 1}{\left(\frac{R_1 R_2}{R_1 + R_2}\right) C_1 s + 1}$	
	Proportional plus rate plus integral controller; lag-lead network	$\frac{e_o}{e_i} = \frac{(R_2 C_2 s + 1)(R_1 C_1 s + 1)}{[(R_1 + R_2) C_2 s + 1] \left(\frac{R_1 R_2}{R_1 + R_2} C_1 s + 1\right)}$ (approximate; see Sec. 1.21)	

network is not loaded. The transfer function for the last network given is approximate and has been obtained by a technique explained in Sec. 1.21.

1.14. The Synthesis of RC Networks. Introduction. The problem of synthesizing a physical network to meet a prescribed transfer function is one that has engaged the attention of a long and distinguished list of investigators. In general, the problem is complicated, and the methods for solving it are correspondingly involved. Fortunately, a solution to the general network-synthesis problem is not normally required in control applications. One reason for this is that the transfer functions that are encountered in control systems are usually quite simple, and ordinarily

they need not be obtained with great precision. Hence, approximate synthesis methods are in almost all cases completely adequate; in fact, one might say that exact synthesis procedures are rarely justified.

A second important reason for the simplification of the synthesis problem is that, as a result of the disadvantages listed in Sec. 1.9, inductances are very rarely used in networks designed for control applications. The restriction of the problem to that of RC network synthesis moves a large number of transfer functions that are theoretically realizable by an unrestricted network into the region of unrealizable functions. It therefore reduces the scope of the synthesis problem considerably. This limitation on the choice of permissible transfer functions is not, however, a very serious disadvantage, since almost all the transfer functions normally required in control systems are found to be of the type realizable with RC networks. It is usually found that the few transfer functions that cannot be synthesized by RC networks alone can be synthesized approximately by RC networks together with amplifying circuits.

1.15. Properties of RC Networks. Before the synthesis of networks is considered, a few remarks concerning some of the properties of RC networks and the physical realizability of transfer functions are in order.

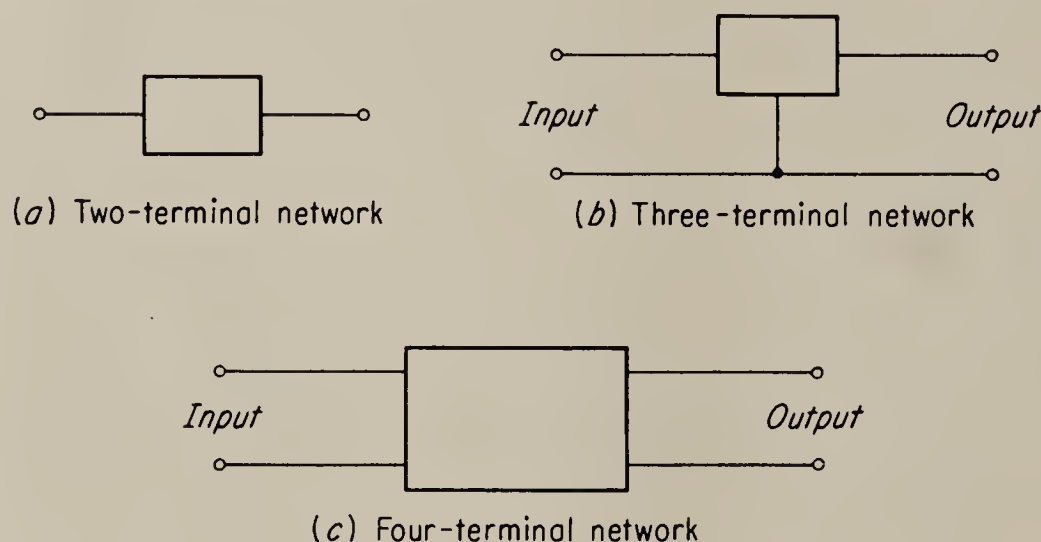


FIG. 1.11. Network types.

The subject of network analysis and synthesis is much too large and complex to be covered adequately in a few paragraphs, and the following statements can serve at best as only a brief summary. For more extensive treatments the reader is referred to standard works¹ on the subject.

Networks may be classified according to the number of terminals by which they are connected to other circuits. Although a network may have as many terminals as it has nodal points, most of the networks in common use are two-, three-, and four-terminal networks, as shown in Fig. 1.11. Note that a three-terminal network is a special type of four-

¹ Bode, "Network Analysis and Feedback Amplifier Design," D. Van Nostrand Company, Inc., Princeton, N.J., 1945. Guillemin, "Communications Networks," John Wiley & Sons, Inc., New York, 1948, vol. 2.

terminal network in which both input and output have a common ground connection. Networks with more than one set of input terminals are occasionally encountered in control systems, particularly in places where a number of signals must be added or subtracted. This type of network will not, however, be discussed in any detail in this chapter. We shall be concerned primarily with three-terminal networks, since this is the type most commonly used in control systems.

A three- or four-terminal network may contain a large number of components, but it is usually possible and convenient to think of it as being made up of a relatively small number of two-terminal networks. For example, the ladder network shown in Fig. 1.12 is made up of the two-terminal networks Z_1, Z_2, Z_3, Z_4 , etc., each one of which may contain a number of resistors and capacitors. The properties of the larger structure are related to those of the two-terminal networks of which it is composed, and it is, therefore, appropriate to consider first the properties of two-terminal networks.

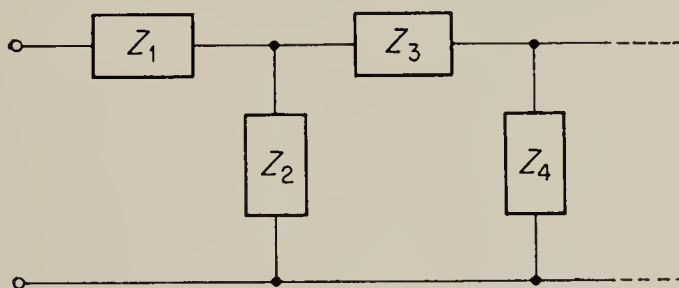


FIG. 1.12. Ladder network.

1.16. Properties of Two-terminal RC Networks. A two-terminal network is characterized by a single function, its *driving-point impedance*, defined as the ratio of the voltage applied to the two terminals to the resulting current. In common with all other network functions, the driving-point impedance is a function of frequency or, more generally, of the complex variable s . The most important property of a driving-point impedance of any passive network made up of lumped parameters such as resistors, capacitors, or inductances is that it can be expressed as the ratio of two polynomials in s as follows:

$$Z = \frac{a_n s^n + a_{n-1} s^{n-1} + \cdots + a_0}{b_m s^m + b_{m-1} s^{m-1} + \cdots + b_0} \quad (1.17)$$

This may be demonstrated by writing all the Kirchhoff mesh equations for the network and solving for the current as a function of the voltage at any two terminals.¹ Equation (1.17) may be written in the equivalent factored form:

$$Z = K \frac{(s - \alpha_1)(s - \alpha_2) \cdots (s - \alpha_n)}{(s - \beta_1)(s - \beta_2) \cdots (s - \beta_m)} \quad (1.18)$$

In accordance with standard convention the α 's and β 's are referred to as the zeros and poles, respectively, of the network function.

Guillemin² has shown that the poles and zeros of RC driving-point

¹ Guillemin, *ibid.*, p. 208.

² *Ibid.*, p. 212.

impedances must be real and negative and that they must obey the separation principle as follows:

$$0 \geq \beta_1 > \alpha_1 > \beta_2 > \alpha_2 > \cdots \quad (1.19)$$

This implies that the zeros and poles must be simple; i.e., factors of the form $(s - \alpha)^n$ cannot occur in the transfer function unless $n = 1$. It follows that an asymptotic representation of an RC driving-point impedance must take the form shown in Fig. 1.13 and that the actual driving-point impedance must satisfy the relation

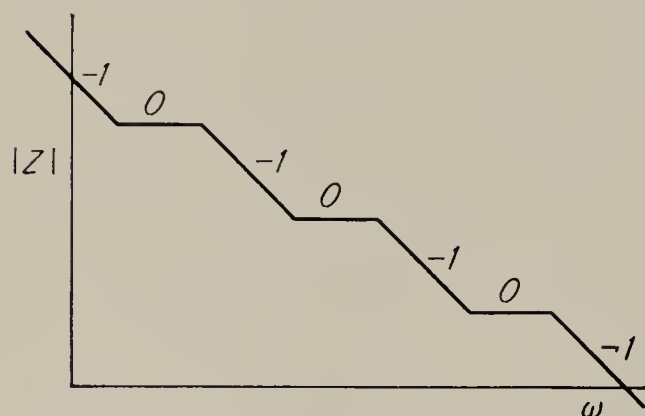


FIG. 1.13. Asymptotic diagram of an RC driving-point impedance.

$$0 \geq \frac{d(\log |Z|)}{d(\log \omega)} \geq -1 \quad (1.20)$$

Bode¹ has shown by means of an argument involving the energy stored or dissipated in a network that the phase angle of an RC driving-point impedance is restricted to negative values between 0 and -90° . This result would also have been suggested by the

well-known relationship between the slope of the asymptotic diagram and the phase angle.

Further implications of Eqs. (1.19) and (1.20) are that the number of poles may be equal to the number of zeros or may exceed the number of zeros by 1. In the former case the driving-point impedance approaches the impedance of a resistance at high frequencies, while in the latter case it approaches that of a capacitance. A driving-point impedance may be infinite or finite, but not zero, at zero frequency, and it may be finite or zero, but not infinite, at infinite frequency.

All of these properties of RC driving-point impedances may be summarized by saying that a network made up of resistors and capacitors will in general have an impedance function that is at all times somewhere between the impedance function of a pure resistance and that of a pure capacitance. The extremes are in general approached only asymptotically. Stated in this way, the conclusions concerning the properties of RC driving-point impedances become almost obvious.

1.17. Transfer Functions Realizable with Four-terminal RC Networks.

Passing now to a consideration of three- and four-terminal networks, we find that such networks are characterized by three network functions, such as, for example, the driving-point impedances at the input and output terminals and the transfer function between input and output voltages. Since our chief interest in this section is the transfer function, we investigate the limitations on the transfer functions realizable by RC

¹ Bode, *op. cit.*, par. 7.11.

networks. In this discussion it is assumed that the driving source supplying the input to the network has zero output impedance and that a load of infinite impedance is connected across the output terminals.

The restrictions on transfer functions obtainable by RC networks are perhaps most simply obtained by noting that any transfer function that is realizable by an RC network may be realized in the form of a symmetrical lattice,¹ or bridge, as shown in Fig.

1.14. In order to find the output voltage, we observe that

$$\hat{e}_o = \hat{e}_2 - \hat{e}_1 \quad (1.21)$$

However
$$\hat{e}_1 = \frac{Z_B}{Z_A + Z_B} \hat{e}_i \quad (1.22)$$

and
$$\hat{e}_2 = \frac{Z_A}{Z_A + Z_B} \hat{e}_i$$

Hence
$$\frac{\hat{e}_o}{\hat{e}_i} = \frac{Z_A - Z_B}{Z_A + Z_B} \quad (1.23)$$

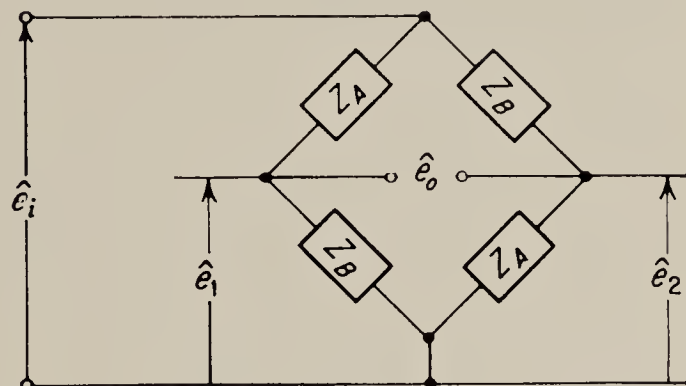


FIG. 1.14. Symmetrical bridge, or lattice, network.

The impedances Z_A , Z_B , and $Z_A + Z_B$ are driving-point impedances and must, therefore, obey the rules governing driving-point impedances determined above. The poles of $Z_A + Z_B$ and the poles of $Z_A - Z_B$ are identical and cancel out of Eq. (1.23); hence the nature of the transfer function is determined completely by the zeros of $Z_A + Z_B$ and $Z_A - Z_B$. It follows immediately that the transfer function must be expressible as the ratio of two polynomials in s . Since the poles of the transfer function are the zeros of the driving-point impedance $Z_A + Z_B$, they must be negative, real, and simple. The zeros of the transfer function, on the other hand, are not similarly restricted and may in general lie anywhere in the complex plane. If the zeros are complex, they must, however, occur in complex-conjugate pairs. This is required, since otherwise the coefficients of the numerator polynomial would be complex numbers, a physical impossibility since these coefficients are combinations of real R 's and C 's. The transfer function cannot have a pole at the origin, since the gain for zero frequency in that case would become infinite. This property may also be deduced from the fact that a driving-point impedance cannot have a zero at the origin [Eq. (1.19)]. Finally the degree of the numerator cannot exceed that of the denominator, since otherwise the gain of the network would become infinite at infinite frequencies. This restriction may again be deduced also from the number of zeros of $Z_A - Z_B$ and of $Z_A + Z_B$.

Before leaving this discussion, we note that, if inductances as well as resistances and capacitances could be used, the only change in the restric-

¹ J. L. Bower and P. F. Ordung, The Synthesis of Resistor-Capacitor Networks, *Proc. IRE*, vol. 38, no. 3, pp. 263-269, March, 1950.

tions on the form of the permissible transfer functions would be in the location of the poles. For a network composed of R 's, L 's, and C 's, the poles may be complex-conjugate and are otherwise restricted only to the left half plane, the latter restriction arising from the fact that the network would otherwise be unstable. Thus, as long as transfer functions having complex poles are not required, RC networks are completely adequate to meet the demand for any realizable transfer function.

1.18. L Sections. Although any transfer function that is realizable by an RC network can be synthesized by a lattice structure, the network most commonly used in control applications is the ladder network shown in Fig. 1.12. This type of network is relatively easy to design, particularly by approximate methods, and despite the fact that it is considerably less versatile than the lattice, it usually meets the requirements of the designer of control systems.

In the following paragraphs we consider an approximate synthesis method based on cascading L-type networks to form a ladder structure. Exact methods of ladder-network synthesis exist but usually result in networks having a considerably lower gain than the approximate design. For details on the design of exact ladder networks and also the design of lattice networks, the reader is referred to the literature.¹

The form of the L section is shown in Fig. 1.15. It is a simple voltage divider, and its transfer function is given by

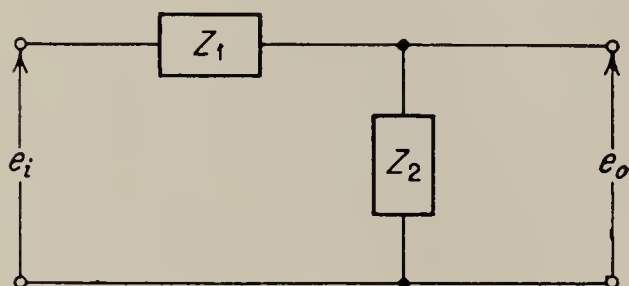


FIG. 1.15 L-section type of RC network.

$$\frac{\hat{e}_o}{\hat{e}_i} = H = \frac{Z_2}{Z_1 + Z_2} \quad (1.24)$$

We make the standard assumptions that the source impedance of the input voltage source is zero and that the load impedance is infinite at all frequencies.

The properties of the network are then easily deduced from the properties of driving-point impedances derived previously. The restrictions on the transfer functions obtainable with L sections may be stated as follows:

1. The poles and zeros must be negative and real.
2. The poles must be simple.
3. There can be no more than two successive zeros (or poles) without an intervening pole (or zero).
4. The number of zeros cannot exceed the number of poles.
5. There cannot be a pole at the origin.

¹J. T. Fleck and P. F. Ordung, Realization of a Transfer Ratio by Means of an R - C Ladder Network, *Proc. IRE*, vol. 39, pp. 1069–1074, 1951. Bower and Ordung, *op. cit.*

6. The maximum rate of increase or decrease of the magnitude of the transfer function cannot exceed the first power of frequency; i.e., the transfer function $H(s)$ must satisfy the inequality

$$-1 \leq \frac{d(\log |H|)}{d(\log \omega)} \leq 1 \quad (1.25)$$

7. The magnitude of the transfer function cannot exceed unity at any frequency, i.e.,

$$|H| \leq 1 \quad (1.26)$$

Equation (1.24) indicates that the poles and zeros of the transfer function are produced by the poles or zeros of Z_2 or $Z_1 + Z_2$. Since these are both driving-point impedances, the poles and zeros must be real and negative, as noted in point 1 of the list above. Poles of the transfer function result only from zeros of $Z_1 + Z_2$, since a pole in Z_2 is also a pole of $Z_1 + Z_2$ and will therefore cancel out. Since zeros of driving-point impedances must be simple, restriction 2 follows. Note, however, that zeros may occur in pairs, because they are produced by either poles of $Z_1 + Z_2$ or zeros of Z_2 . Restriction 3 follows from the separation property of the poles and zeros of driving-point impedances. Thus suppose that H has a pole due to a zero of $Z_1 + Z_2$. The next singularity of $Z_1 + Z_2$ must be a pole; however, if this is a pole of Z_2 , no singularity in H results. $Z_1 + Z_2$ may now again have a zero, resulting in two successive poles in H . The next singularity in H must now, however, be caused by either a pole in Z_1 or a zero in Z_2 . Either of these singularities results in a zero in H . Thus there can be no more than two successive poles. A similar argument may be used to show that no more than two successive zeros may occur. Restrictions 4 and 5 are restrictions on RC -network transfer functions in general and need not be discussed again. Restriction 6 is related to inequality (1.20). The maximum rate of decrease of the transfer function occurs when the numerator decreases at its maximum rate and when the denominator is constant; the maximum rate of increase occurs when the numerator is constant and when the denominator increases at its maximum rate. Thus restriction 6 follows. Restriction 7 is almost obvious, since the network contains no amplifier, but it may be shown more rigorously to be a consequence of the limitation on the phase angles of Z_1 and Z_2 . Thus, suppose Eq. (1.24) to be rewritten as follows:

$$H = \frac{Z_2}{Z_1 + Z_2} = \frac{1}{(Z_1/Z_2) + 1} \quad (1.27)$$

Since the phase angles of Z_2 and Z_1 are restricted to the fourth quadrant,

we may write

$$-90^\circ \leq \angle \frac{Z_2}{Z_1} \leq 90^\circ \quad (1.28)$$

Hence

$$\frac{Z_1}{Z_2} + 1 \geq 1 \quad (1.29)$$

and restriction 7 is proved.

1.19. Cascading of L Sections. Restrictions 2, 3, and 6 may be overcome by cascading several L sections to form a ladder network of the form

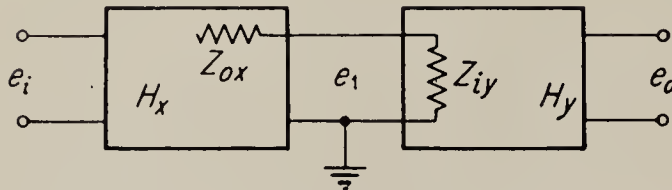


FIG. 1.16. Cascaded networks.

shown in Fig. 1.12. If the input impedance of the second L section is sufficiently high relative to the output impedance of the first section, so that the second section presents negligible load on the first, then the

transfer function through both sections is approximately equal to the product of the transfer functions of the individual sections. This is easily demonstrated.

Suppose a network having a transfer function H_x and an output impedance Z_{ox} drives a network having a transfer function H_y and an input impedance Z_{iy} (see Fig. 1.16). When the second network is not connected, the output of the first network, e_1 , is given by

$$\hat{e}_1 = \hat{e}_i H_x \quad (1.30)$$

With the second network connected, however, the voltage e_1 becomes

$$\hat{e}_1 = \hat{e}_i H_x \frac{Z_{iy}}{Z_{ox} + Z_{iy}} \quad (1.31)$$

Therefore, the output voltage from the second network is given by

$$\hat{e}_o = \hat{e}_1 H_y = \hat{e}_i H_x H_y \frac{Z_{iy}}{Z_{ox} + Z_{iy}} \quad (1.32)$$

or

$$\frac{\hat{e}_o}{\hat{e}_i} = \frac{H_x H_y}{1 + (Z_{ox}/Z_{iy})} \quad (1.33)$$

It is clear that, if

$$\left| \frac{Z_{ox}}{Z_{iy}} \right| \ll 1 \quad (1.34)$$

then

$$\frac{\hat{e}_o}{\hat{e}_i} \approx H_x H_y \quad (1.35)$$

The maximum permissible value of $|Z_{ox}/Z_{iy}|$ depends on the accuracy desired, but a commonly accepted maximum is of the order of 0.1. Normally only a very rough estimate of $|Z_{ox}/Z_{iy}|$ is required. If more than two L sections are to be cascaded, the above argument is easily extended, and inequality (1.34) must be met at each network junction.

In computing the input and output impedances of an L section, it is convenient to note that the input impedance of network H_y is given by

$$Z_{iy} = Z_{1y} + Z_{2y} = \frac{Z_{2y}}{H_y} \quad (1.36)$$

while the output impedance of network H_x is given by

$$Z_{ox} = \frac{Z_{1x}Z_{2x}}{Z_{1x} + Z_{2x}} = Z_{1x}H_x \quad (1.37)$$

Hence

$$\frac{Z_{ox}}{Z_{iy}} = \frac{Z_{1x}H_xH_y}{Z_{2y}} \quad (1.38)$$

When three or four L sections must be cascaded, the application of inequality (1.34) may require unrealistic component values, since the inequality requires the impedance level of each succeeding section to be several times higher than the preceding one. Impedance levels larger than 1 megohm are usually undesirable, with their probability of excessive noise pickup. Hence isolating amplifiers or cathode followers (see Sec. 2.5) are quite frequently used to prevent loading between sections in the more extensive networks.

If the number of cascaded L sections is N , then restrictions 2, 3, and 6 of the list in Sec. 1.18 are amended as follows:

2*a*. The poles of the transfer function may be approximately N -fold. Note that the requirement that the poles must be simple was made in general for all RC networks. Hence, even with N cascaded sections it is not possible to have N exactly coinciding poles; however by making Z_{ox}/Z_{iy} sufficiently small, a number of poles may be brought as close together as is desired.

3*a*. There may be $2N$ successive zeros (or poles) without an intervening pole (or zero).

6*a*. The maximum rate of increase or decrease of the magnitude of the transfer function cannot exceed the N th power of frequency:

$$-N \leq \frac{d(\log H)}{d(\log \omega)} \leq N \quad (1.39)$$

1.20. Approximate Ladder-network Synthesis. In synthesizing a network to obtain a prescribed transfer function, it is first necessary to ascertain that the desired transfer function meets all the requirements listed above. When this has been done, the number of L sections required is determined by examination of the amended restrictions 2*a*, 3*a*, or 6*a*. Thus, for instance, if a transfer function having an asymptotic gain diagram with a -3 slope is desired, at least three L sections would be needed. The apportionment of poles and zeros among the various L sections is to some extent arbitrary, and any order of cascading the parts is usually

permissible as long as inequality (1.34) is properly satisfied. The specific design chosen from this range of possibilities should be that which minimizes the size of the larger capacitors and makes inequality (1.34) easy to meet. In order to accomplish the former, it is usually best to place the sections having zeros and poles of smallest magnitudes in the last position of the cascade.

Examples will be worked out in a later section to provide further clarification of the design procedure. Before a design can be completed, however, a method must be available for the synthesis of the individual L section. We present here three methods, one approximate and the others exact.

1.21. Approximate Synthesis of L Sections. The following approximate synthesis method is based on the idea that a capacitance connected in a circuit with resistances acts as an open circuit at very low frequencies and as a short circuit at very high frequencies. For each capacitor in the circuit, therefore, there exists only a relatively small frequency range for which its impedance is comparable to that of the resistances surrounding it. In this frequency range we consider the capacitor as being *active*. For other frequencies, where the capacitor represents either an open circuit or a short circuit relative to the resistances connected to it, we consider it *inactive*, since it has negligible effect on the circuit.

The synthesis method is approximate rather than exact because it assumes that the region of activity of a given capacitor is sharply bounded. Under this assumption a capacitor is either active or it is inactive, and no intermediate conditions are possible. This implies that the change from activity to inactivity takes place at distinct boundary points on the frequency scale. In order to locate these boundary points, suppose first that a driving-point impedance contains, at a particular frequency, no active capacitors. The impedance is then made up only of resistances, and the slope of the asymptotic impedance diagram (as in Fig. 1.13) must be zero. When a capacitor becomes active, it changes from an open circuit to a short circuit as the frequency increases. The presence of an active capacitor, therefore, causes a reduction in impedance with frequency as shown by the segments of -1 slope in Fig. 1.13. It is most natural to take the break points of the asymptotic impedance diagram as the boundary points defining the activity of a capacitor. In accordance with the usual convention, the frequencies of the break points are equal in magnitude to the poles and zeros of the impedance. Hence, if it is assumed that there is never more than one active capacitor in an impedance, we may think of this capacitor as being responsible for the pole and the zero defining its zone of activity.

A further important point in the synthesis procedure is that there must never be more than one active capacitor in the entire L section being

synthesized. Although this is not necessarily true for L sections in general, we are certainly free to design an L section so that this stipulation is met. We permit the limiting case where one capacitor becomes active at the same point at which another one becomes inactive, since otherwise the synthesis of transfer functions having two coincident zeros would not be possible.

Consider now the general two-terminal impedance Z , containing only one capacitor (see Fig. 1.17). At frequencies low enough for the capacitor to be inactive, the impedance is effectively a resistance R . At intermediate frequencies the capacitor becomes active and the impedance decreases, and at high frequencies the capacitor is again inactive. Hence over the entire frequency range the impedance may be expressed by

$$Z = R \frac{T_1 s + 1}{T_2 s + 1} \quad (1.40)$$

where

$$T_2 > T_1$$

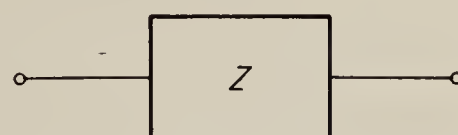


FIG. 1.17. Impedance containing one capacitor.

If a voltage source E is connected across the impedance, the resulting current is

$$\hat{I} = \frac{\hat{E}}{Z} = \frac{\hat{E}(T_2 s + 1)}{R(T_1 s + 1)} \quad (1.41)$$

Rearranging gives

$$R(T_1 s + 1)\hat{I} = (T_2 s + 1)\hat{E} \quad (1.42)$$

Equation (1.42) may be thought of as the transform of the differential equation

$$RT_1 \frac{dI}{dt} + RI = T_2 \frac{dE}{dt} + E \quad (1.43)$$

Suppose now that the impedance is open-circuited. Then the current I is zero, and Eq. (1.43) becomes

$$T_2 \frac{dE}{dt} + E = 0 \quad (1.44)$$

Assuming that the voltage has some initial value E_0 , the solution for E as a function of time is

$$E = E_0 e^{-t/T_2} \quad (1.45)$$

Similarly, we may solve for the current that flows when the two terminals of the impedance are short-circuited together. Under this condition $E = 0$, and Eq. (1.43) becomes

$$T_1 \frac{dI}{dt} + I = 0 \quad (1.46)$$

or

$$I = I_0 e^{-t/T_1} \quad (1.47)$$

From Eq. (1.40) we see that $1/T_1$ and $1/T_2$ are, respectively, the zero and the pole of Z . Thus we arrive at the following conclusion:

When a driving-point impedance contains only one capacitor, it has only one pole and one zero. The pole is the reciprocal of the discharge time constant obtained on the open-circuited impedance, and the zero is the reciprocal of the discharge time constant of the impedance when it is short-circuited.

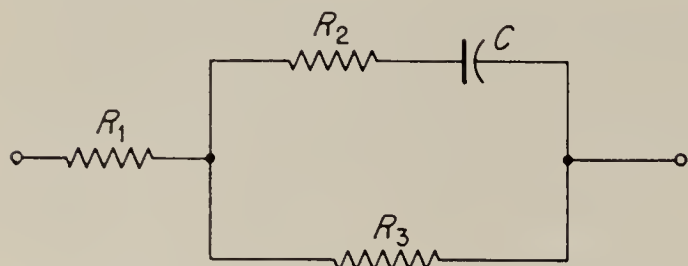


FIG. 1.18. RC impedance.

This rule enables us to write down the driving-point-impedance functions by inspection of the circuit. Thus, consider the impedance shown in Fig. 1.18. At low frequencies the capacitor is an open circuit; hence

the resistance is $R_1 + R_3$. The pole is the reciprocal of $(R_2 + R_3)C$. The zero is the reciprocal of $[R_1R_3/(R_1 + R_3) + R_2]C$. Hence the impedance function is

$$Z = (R_1 + R_3) \frac{[R_2 + (R_1R_3)/(R_1 + R_3)]Cs + 1}{(R_2 + R_3)Cs + 1} \quad (1.48)$$

In an L-section network like that shown in Fig. 1.15, a pole in the transfer function is produced only by a zero of $Z_1 + Z_2$ that is not simultaneously a zero of Z_2 . Hence for a network containing only one active capacitor a pole in the transfer function is always the reciprocal of a discharge time constant of the entire short-circuited L section. Zeros in the transfer function are produced by zeros of Z_2 or poles of Z_1 , since a pole of Z_1 corresponds to a pole of $Z_1 + Z_2$. Thus zeros are given either by the discharge time constant of Z_2 short-circuited, or of Z_1 open-circuited.

To synthesize an L-section network, we first draw up the network qualitatively according to the following rules:

1. If the asymptotic representation of the transfer function has a zero slope, the network has no active capacitors in it.
2. If the asymptotic representation of the transfer function has a $+1$ slope, a capacitor in Z_1 is active, and if it has a -1 slope, a capacitor in Z_2 is active.
3. A zero in the transfer function at zero frequency requires a series capacitor, not bridged by any resistance in Z_1 . A zero at infinite frequency requires a shunt capacitor across Z_2 .

The procedure is best illustrated by means of a few typical examples. Consider the transfer function whose asymptotic diagram is shown in Fig. 1.19. The network must contain two capacitors, one active in the interval $\omega_1 \leq \omega \leq \omega_2$ and the other active in the interval $\omega_3 \leq \omega \leq \omega_4$. We consider these two intervals separately. A network to synthesize the transfer function to the left of ω_3 is shown in Fig. 1.20a. Note that at low

frequencies the capacitor acts as an open circuit; hence the network gain is unity. At intermediate frequencies the capacitor becomes active and reduces the gain, and at high frequencies the capacitor is short-circuited, and the network gain remains constant at a reduced level. For frequencies above ω_3 the network gain must increase. By rule 2 this requires a capacitor in Z_1 . If the network gain is to be less than unity at high frequencies, however, the capacitor must short out only part of R_1 . Thus the complete network shown in Fig. 1.20*b* results.

To complete the design quantitatively, we note that the break at ω_1 represents a pole in the transfer function of magnitude ω_1 . This pole is caused by a zero of the input driving-point impedance of the circuit of

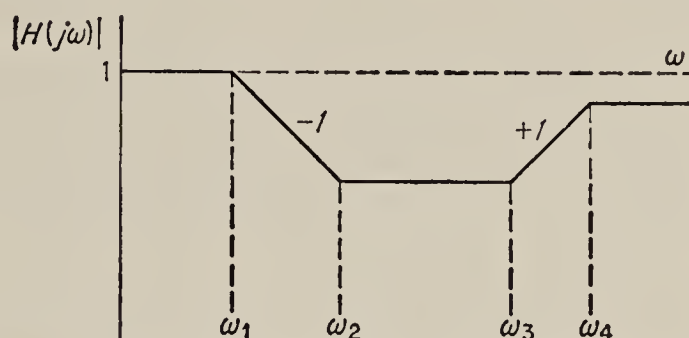


FIG. 1.19. Example of transfer function.

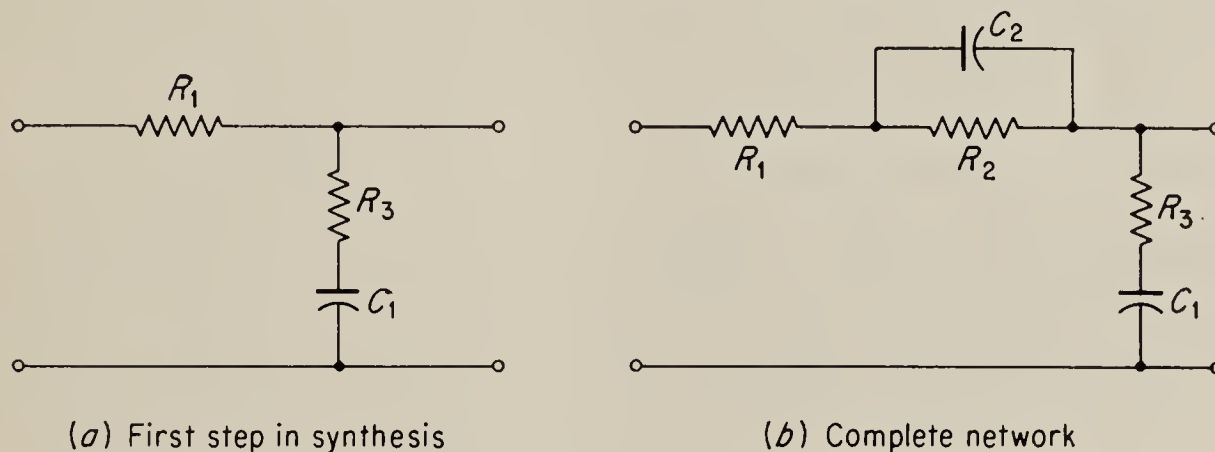


FIG. 1.20. Synthesis of the transfer function of Fig. 1.19.

Fig. 1.20*a* and is therefore the reciprocal of the discharge time constant of that circuit when the input is short-circuited. Thus

$$\frac{1}{\omega_1} = (R_1 + R_2 + R_3)C_1 \quad (1.49)$$

The break at ω_2 represents a zero of Z_2 which is equal to the reciprocal of the discharge time constant of Z_2 when its two ends are short-circuited; i.e.,

$$\frac{1}{\omega_2} = R_3C_1 \quad (1.50)$$

The next zero, corresponding to ω_3 , is due to a pole of Z_1 . It is, therefore, the reciprocal of the discharge time constant of Z_1 when it is open-circuited;

$$\frac{1}{\omega_3} = R_2C_2 \quad (1.51)$$

Finally, $1/\omega_4$ must be equal to the discharge time constant of the entire

network with the input terminals connected together and with C_1 replaced by a short circuit.

$$\frac{1}{\omega_4} = \frac{C_2 R_2 (R_1 + R_3)}{R_1 + R_2 + R_3} \quad (1.52)$$

Although other equations may be written about the circuit of Fig. 1.20b, it will be found that they are not independent from the four given above. Thus we have only four independent equations. There are, however, five unknown parameters in the circuit: two capacitors and three resistors. This situation is encountered in all methods of transfer-function synthesis, since the impedance level of the network is not specified by the process. The value of any one component of the network may therefore be set arbitrarily, depending on the requirements of the remainder of the circuit. Thus, suppose the circuit to be driving the grid of a vacuum tube. Then the maximum resistance between grid and ground might be specified as less than 1 megohm, so that we would have the additional requirement that $R_2 + R_1 \leq 10^6$ ohms. If then we arbitrarily let $R_1 + R_2 = \frac{1}{2}$ megohm,

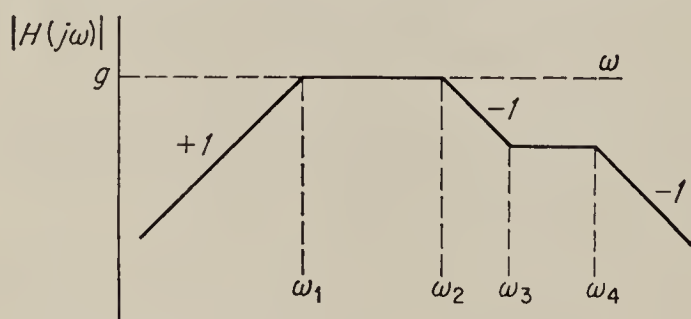


FIG. 1.21. Network transfer function.

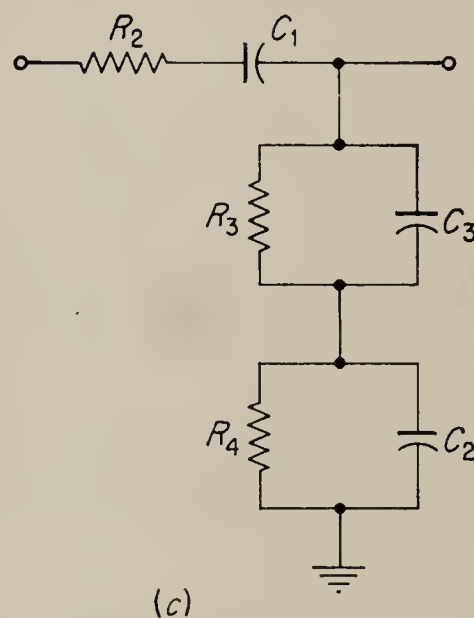
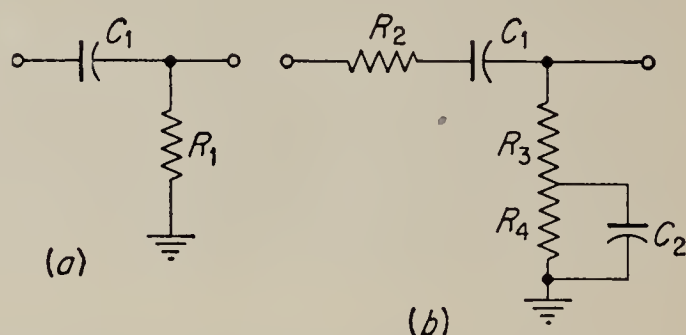


FIG. 1.22. Steps in the design (see text).

we have five equations, which may be solved to find the values of the five parameters. Since most of the equations are, in general, quite simple, like Eqs. (1.50) and (1.51), it is usually possible by a judicious choice of the arbitrary parameter to simplify the solution very considerably. Thus, in the foregoing example, if we assign a value to C_1 , Eq. (1.50) may be solved immediately for R_3 , and Eq. (1.49) for $R_1 + R_2 + R_3$. Then division of Eqs. (1.51) and (1.52) yields an equation having only R_1 as the unknown, so that all the resistances may be obtained easily. C_2 is then obtained from Eq. (1.51).

As a second example, consider the transfer function shown in Fig. 1.21 which has a zero both for zero and for infinite frequencies. The $+1$ slope extending down to zero frequency requires a series capacitor in Z_1 by rule

3. Hence the transfer function to the left of ω_2 is synthesized by the network shown in Fig. 1.22a. At the frequency ω_1 the capacitor C_1 becomes inactive. In order to obtain the -1 slope between ω_2 and ω_3 , a capacitor in Z_2 must become active in this interval; however, unless there is a series resistance in Z_1 , such a capacitor cannot have any effect on the transfer function. Thus the network must take the form shown in Fig. 1.22b. The final break at ω_4 must be produced by another capacitor in Z_2 , so that the network of Fig. 1.22c follows. Note that a capacitance is shunted across Z_2 as required by rule 3.

To evaluate the R 's and C 's, we proceed as in the previous example and obtain the following equations:

$$\frac{1}{\omega_1} = (R_2 + R_3 + R_4)C_1 \quad (C_2 \text{ and } C_3 \text{ are open circuits}) \quad (1.53)$$

$$\frac{1}{\omega_2} = \frac{C_2 R_4 (R_3 + R_2)}{R_4 + R_2 + R_3} \quad \begin{array}{l} (C_1 \text{ is short-circuited and} \\ C_3 \text{ is open-circuited}) \end{array} \quad (1.54)$$

$$\frac{1}{\omega_3} = R_4 C_2 \quad \begin{array}{l} (C_1 \text{ is short-circuited and} \\ C_3 \text{ is open-circuited}) \end{array} \quad (1.55)$$

$$\frac{1}{\omega_4} = \frac{C_3 R_2 R_3}{R_2 + R_3} \quad \begin{array}{l} (C_1 \text{ and } C_2 \text{ are short-} \\ \text{circuited}) \end{array} \quad (1.56)$$

We have four equations but six unknowns. One more equation may be obtained, as in the last example, by assigning an arbitrary value to one of the components, but this still leaves one additional equation to be determined. This additional equation is required to specify the gain level of the network. Thus, suppose that the gain of the transfer function between the frequencies ω_1 and ω_2 is set at g , where $g < 1$. Then since C_1 is assumed short-circuited and C_2 and C_3 open-circuited in this frequency interval, we have

$$g = \frac{R_3 + R_4}{R_2 + R_3 + R_4} \quad (1.57)$$

Thus we have the necessary six equations, which may now be solved for the six components.

It should be noted that the question of gain level did not appear in the first example because the level was tacitly assumed to be unity at frequencies below ω_1 . Had a level of less than unity been desired, then an additional resistor (probably across C_1 in Fig. 1.20b) would have been placed in the circuit. An additional equation would then have been needed to specify this resistance.

Normally one is interested in as high a gain level as possible, since low gain must be made up with amplifying circuits. The question naturally arises, therefore, as to how large g may be made. For a transfer function of the form shown in Fig. 1.19 the answer is obvious: the maximum gain is unity. However, for network functions of the type shown in Fig. 1.21,

the answer is not obvious, and it cannot be obtained from the approximate synthesis procedure described here. All that can be said at this point is that $g < 1$, since, by Eq. (1.57), $R_2 = 0$ if $g = 1$.

The essential difference between the transfer functions of Figs. 1.19 and 1.21 is that the latter contains two successive poles, while the former does not. It will be shown in connection with the exact synthesis method to

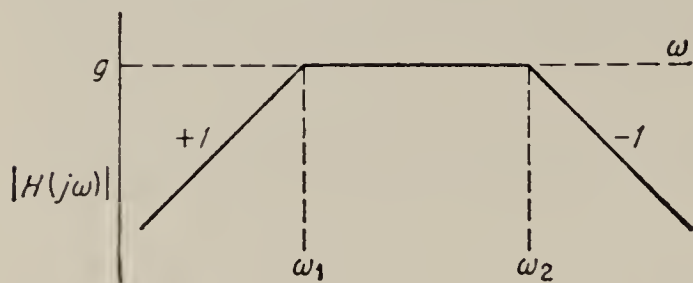


FIG. 1.23. Transfer function with two successive poles.

be described subsequently (Sec. 1.23) that this question of gain level arises generally only when two successive poles are specified for a transfer function. An approximate idea of the maximum value of g permissible in such cases may, therefore, be obtained by considering a transfer function having only two poles. The asymptotic diagram of such a transfer function must have the form shown in Fig. 1.23, and an analytic expression of H is

$$H = \frac{Z_2}{Z_1 + Z_2} = \frac{g\omega_2 s}{(s + \omega_1)(s + \omega_2)} \quad (1.58)$$

This may be rewritten as

$$\frac{Z_1 + Z_2}{Z_2} = 1 + \frac{Z_1}{Z_2} = \frac{(s + \omega_1)(s + \omega_2)}{g\omega_2 s} \quad (1.59)$$

so that

$$\frac{Z_1}{Z_2} = \frac{s^2 + (\omega_1 + \omega_2)s + \omega_1\omega_2 - g\omega_2 s}{g\omega_2 s} \quad (1.60)$$

The roots of the numerator must be real and negative. By a short algebraic manipulation it may be shown that this requirement leads to

$$g \leq \left(1 - \sqrt{\frac{\omega_1}{\omega_2}}\right)^2 \quad (1.61)$$

A plot of this function is shown in Fig. 1.24. Note that for $\omega_2 = \omega_1$ the gain is zero; this means that a network having two coincident poles cannot be synthesized. This was, of course, proved previously by use of a different argument. However, even if the poles are separated by a factor of 10, the maximum gain is only 0.45. This indicates that, unless they are very far apart, successive poles should be avoided as much as possible. When a number of L sections must be cascaded for other reasons, it is therefore often best

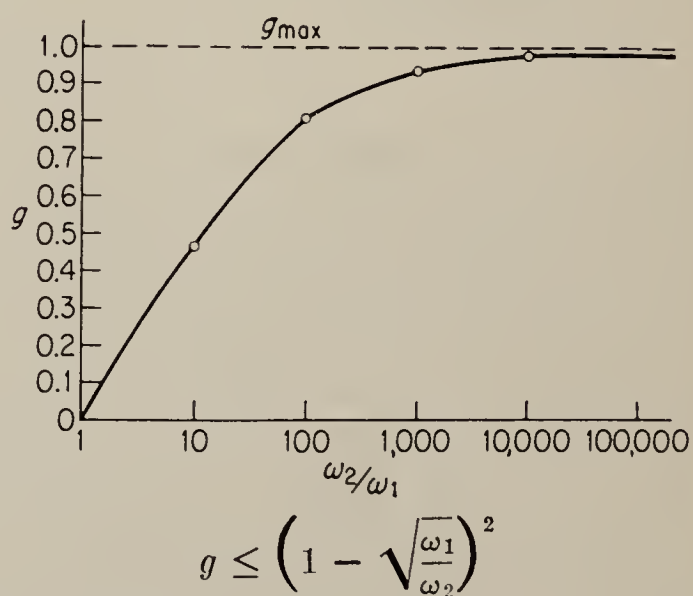


FIG. 1.24. Maximum gain for network having transfer function of Fig. 1.23.

$$g \leq \left(1 - \sqrt{\frac{\omega_1}{\omega_2}}\right)^2$$

to locate successive poles in different L sections. Finally, it should be noted that no difficulty is experienced in synthesizing a transfer function of the form shown in Fig. 1.25 by a single L section. Here the gain in the region having two poles is assumed to be considerably below the limiting value given by Eq. (1.61), and the approximate method should give good results.

Of considerable interest in connection with any approximate method is the question of how good the approximation is. While in the network-synthesis problem no general answer applicable in all cases can be given, a number of points can be made as follows:

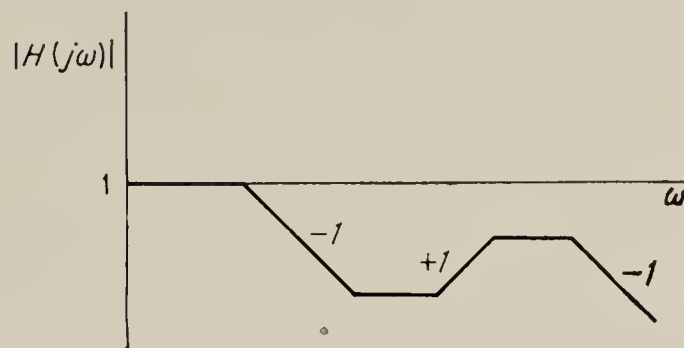


FIG. 1.25. Transfer function easily realizable by single L section.

1. If the network contains only one capacitor, the approximate and exact methods give identical results.

2. Any pole or zero of the transfer function that is produced by a pole or zero of an impedance having only one capacitor is given exactly by the approximate method. Thus in the network of Fig. 1.20b the zeros ω_2 and ω_3 are exact, since the first is due to a zero of Z_2 and the second is due to a pole of Z_1 , and both of these impedances have only one capacitor.

3. Any zero of the transfer function produced by a pole of Z_1 is exact. This is due to the fact that the approximate method is inaccurate only in locating the zeros of driving-point impedances.

4. In general, the approximate method gains in accuracy as the zones of activity of the various capacitors contained in the impedances are moved farther apart on the frequency scale. It is clear that under this condition the approximation that a capacitor is either open- or short-circuited outside its active zone becomes better and better.

1.22. An Exact Synthesis Method of L Sections for Simple Transfer Functions. For the relatively simple network functions that are used in the great majority of control systems, the approximate synthesis method discussed in the previous section is quite easily made exact. The procedure is best illustrated by an example. Thus consider the transfer function diagrammed in Fig. 1.19 above. By means of the qualitative reasoning discussed in connection with this figure, we obtain as before the general network shown in Fig. 1.20b. The exact transfer function of this network is obtained by treating it as a voltage divider:

$$\begin{aligned}
 H &= \frac{R_3 + (1/C_1s)}{R_3 + (1/C_1s) + R_1 + [R_2/(R_2C_2s + 1)]} \\
 &= \frac{(R_3C_1s + 1)(R_2C_2s + 1)}{(R_1 + R_3)R_2C_2C_1s^2 + (R_1C_1 + R_3C_1 + R_2C_1 + R_2C_2)s + 1} \quad (1.62)
 \end{aligned}$$

From Fig. 1.19 the transfer function is given by

$$H = \frac{[(s/\omega_2) + 1][(s/\omega_3) + 1]}{(s^2/\omega_1\omega_4) + [(1/\omega_1) + (1/\omega_4)]s + 1} \quad (1.63)$$

By direct comparison we obtain

$$\begin{aligned} R_2C_2 &= \frac{1}{\omega_3} \\ R_3C_1 &= \frac{1}{\omega_2} \end{aligned} \quad (1.64)$$

Substituting these values in Eq. (1.62) and equating Eqs. (1.62) and (1.63), we obtain

$$\begin{aligned} \frac{s^2}{\omega_1\omega_4} + \left(\frac{1}{\omega_1} + \frac{1}{\omega_4}\right)s + 1 &= \frac{1}{\omega_3} \left(\frac{1}{\omega_2} + R_1C_1\right)s^2 \\ &+ \left(R_1C_1 + R_2C_1 + \frac{1}{\omega_2} + \frac{1}{\omega_3}\right)s + 1 \end{aligned} \quad (1.65)$$

By equating the coefficients of the s^2 term, we obtain

$$R_1C_1 = \frac{\omega_3}{\omega_1\omega_4} - \frac{1}{\omega_2} \quad (1.66)$$

Substituting this value for R_1C_1 in Eq. (1.65) and equating the coefficients of s , we obtain finally

$$R_2C_1 = \frac{1}{\omega_1} + \frac{1}{\omega_4} - \frac{\omega_3}{\omega_1\omega_4} - \frac{1}{\omega_3} \quad (1.67)$$

Equations (1.64), (1.66), and (1.67), along with the specification of one of the R 's or C 's to provide the desired impedance level for the network, constitute a set of five equations, from which the five network parameters are easily obtained.

This method may be used for other simple networks having only a few poles and zeros. As the networks become more complicated, considerable ingenuity may have to be exercised in particular cases to make the substitutions indicated above. Also the method does not optimize the gain level for transfer functions having two successive poles, like that shown in Fig. 1.21. Such optimization would have to be a trial-and-error process with larger and larger values of gain assumed until an unrealizable situation (a negative value for an RC product, for instance) appears. Despite its shortcomings, the method has considerable merit where applicable, since it is considerably simpler than the general exact synthesis method to be described in the next section.

1.23. The Exact Synthesis of L Sections. General Method.¹ The L section considered in this section is that shown in Fig. 1.15. Suppose

¹ J. L. Bower, P. F. Ordnung, and J. T. Fleck, "The Synthesis of Resistor-Capacitor Networks," Yale Univ. Dept. Elec. Eng. Tech. Rept., August, 1948, pp. 21-24.

that the desired transfer function meeting all the restrictions for L sections discussed above has the form

$$H = \frac{Z_2}{Z_1 + Z_2} = h \frac{(s + \alpha_1)(s + \alpha_2)(s + \alpha_3)}{(s + \beta_1)(s + \beta_2)(s + \beta_3)} \quad (1.68)$$

In this expression h is a constant gain multiplier which is to be made as large as possible. Equation (1.68) may be rewritten as follows:

$$\frac{Z_1 + Z_2}{Z_2} = 1 + \frac{Z_1}{Z_2} = \frac{1}{h} \frac{(s + \beta_1)(s + \beta_2)(s + \beta_3)}{(s + \alpha_1)(s + \alpha_2)(s + \alpha_3)} \quad (1.69)$$

Defining

$$F(s) \triangleq \frac{(s + \beta_1)(s + \beta_2)(s + \beta_3)}{(s + \alpha_1)(s + \alpha_2)(s + \alpha_3)} \quad (1.70)$$

we find that

$$\frac{Z_1}{Z_2} = \frac{1}{h} [F(s) - h] \quad (1.71)$$

From this equation it is clear that poles of Z_1 or zeros of Z_2 result from poles of $F(s)$, while zeros in Z_1 or poles in Z_2 result whenever $F(s) - h = 0$.

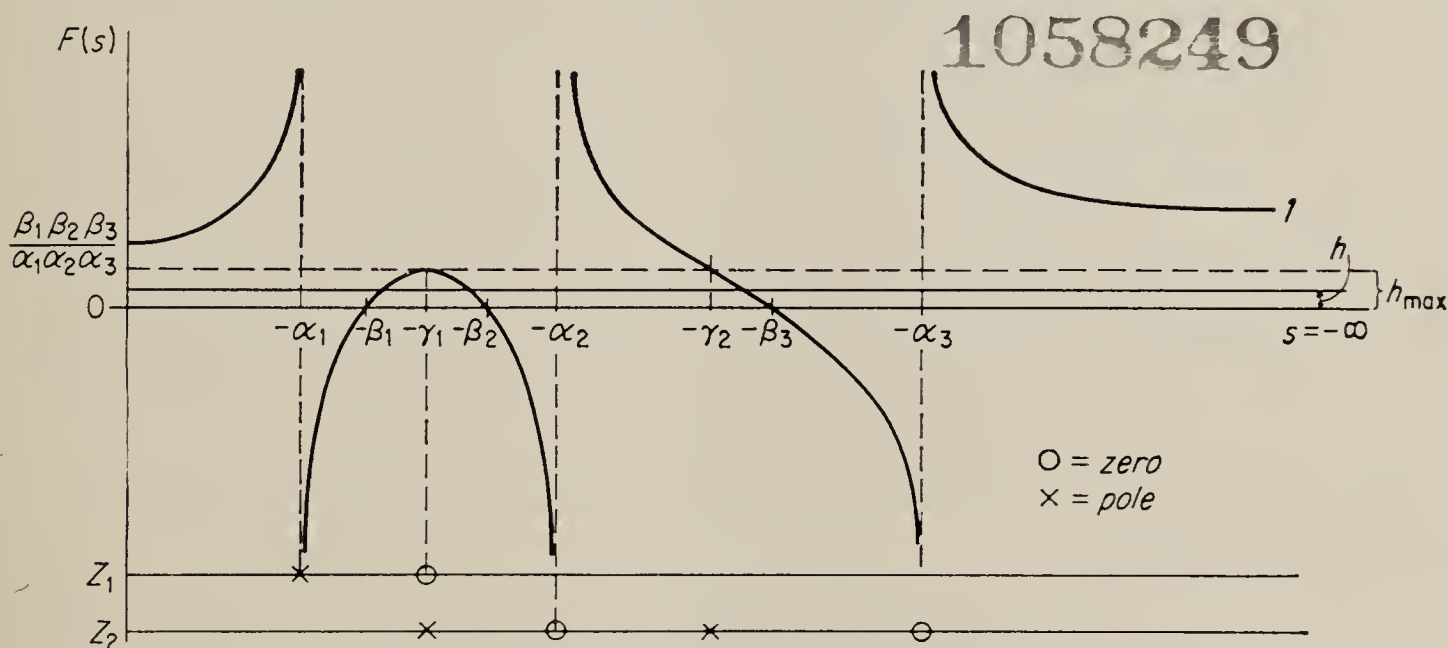


FIG. 1.26. $F(s)$ versus s on a distorted abscissa that includes both $s = 0$ and $s = \infty$.

The zeros of $F(s) - h$ are best found graphically, particularly if $F(s)$ is of higher than the second degree. A convenient method is to make a plot of $F(s)$ for negative real values of s . A typical plot is shown in Fig. 1.26 for the case $0 < \alpha_1 < \beta_1 < \beta_2 < \alpha_2 < \beta_3 < \alpha_3$. The abscissa of the plot is significant only in certain regions and may be made approximately logarithmic in s with the ends distorted to include zero and infinity. Furthermore $F(s)$ does not need to be accurately plotted, and a rough sketch suffices.

A horizontal line representing an arbitrary value of h is drawn on this sketch, and each intersection of this line with the curve for $F(s)$ is therefore a zero of $F(s) - h$. In order to determine the maximum value of h ,

we note that, in the example given, $F(s) - h = 0$ is a cubic equation and therefore has three roots. For these roots to be negative and real the line representing h must intersect $F(s)$ three times for negative real s . Hence the maximum value of h is seen to be equal to $F(-\gamma_1)$, the maximum of $F(s)$ occurring between β_1 and β_2 . A horizontal line corresponding to this value of h is shown dotted in Fig. 1.26. In general, a similar limitation on the maximum value of h can always be found.

The example given illustrates again the effect of successive poles in the transfer function on the gain level of the network. Note that the limit on h_{\max} is set by the peak in the $F(s)$ function occurring between the two poles of H , that is, between β_1 and β_2 . It should be clear from the construction that for maximum gain these poles should be separated as far as possible.

The exact values of h_{\max} and the zeros γ_1 and γ_2 must be known before the synthesis can be made. These values are best found by a simple trial-and-error procedure, as will be shown later in an example.

The distribution of zeros and poles to the two impedances Z_1 and Z_2 is now made as indicated by the o 's and x 's at the bottom of Fig. 1.26. Starting at the left, the pole at x_1 must be assigned to Z_1 as a pole, since the smallest singularity in any impedance must be a pole. At the point $s = \gamma_1$ there are two coincident zeros of $F(s) - h$; these are produced by a zero in Z_1 and a pole in Z_2 . For the three remaining singularities there is a variety of possible arrangements. It is usually desirable to keep the total number of poles in the two impedances as small as possible, since each pole calls for a capacitor (see Sec. 1.24). The arrangement shown in Fig. 1.26 results in a design calling for a minimum number of capacitors, but it is not the only arrangement that does so.

The construction up to this point has now given us the impedances Z_1 and Z_2 in the form

$$Z_1 = K_1 \frac{s + \gamma_1}{s + \alpha_1} \quad (1.72)$$

$$Z_2 = K_2 \frac{(s + \alpha_2)(s + \alpha_3)}{(s + \gamma_1)(s + \gamma_2)} \quad (1.73)$$

The α 's and γ 's are known, but K_1 and K_2 are not yet determined. The ratio K_1/K_2 is found most easily by using Eq. (1.71) evaluated at a convenient value of s . Thus, for $s = -\infty$, Eq. (1.71) becomes

$$\left. \frac{Z_1}{Z_2} \right|_{s=-\infty} = \frac{K_1}{K_2} = \frac{1}{h} (1 - h) \quad (1.74)$$

One unknown parameter now remains. As noted previously, this arises from the fact that in synthesizing a transfer function we do not consider the impedance level of the resulting network. Thus the single unknown

parameter that remains is arbitrary and permits adjustment of the impedance level.

1.24. Exact Synthesis of RC Driving-point Impedances. The final problem remaining is the synthesis of driving-point-impedance functions such as those obtained in Eqs. (1.72) and (1.73) in terms of R 's and C 's. This may be done in a variety of ways,¹ two of which are given here.

In the first method we make a partial-fraction expansion of the impedance function. Thus, for the impedance given in Eq. (1.73) we have:

$$Z_2 = K_2 + \frac{(\alpha_2 + \alpha_3 - \gamma_1 - \gamma_2)s + \alpha_2\alpha_3 - \gamma_1\gamma_2}{(s + \gamma_1)(s + \gamma_2)} \quad (1.75)$$

$$= K_2 + \frac{K_3}{s + \gamma_1} + \frac{K_4}{s + \gamma_2} \quad (1.76)$$

Note that the partial-fraction expansion is valid only for proper fractions; hence the impedance function of Eq. (1.73) must first be converted by division into the sum of a constant and a proper fraction, as shown in Eq. (1.75).

The constants K_3 and K_4 are found by standard techniques. Equation (1.76) suggests that Z_2 consists of three components in series. The first component may be identified as a resistance K_2 , while the other two both consist of a parallel combination of a resistor and a capacitor, as shown in Fig. 1.27. This

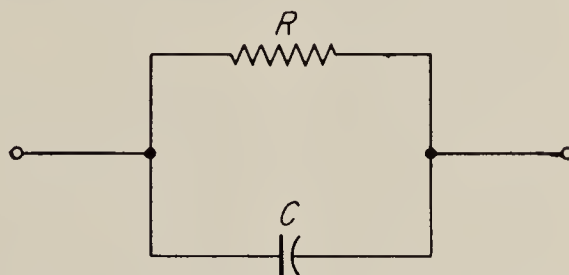


FIG. 1.27. Resistor and capacitor in parallel.

may be shown by evaluating the impedance of this combination:

$$Z = \frac{R/Cs}{R + (1/Cs)} = \frac{1/C}{s + (1/RC)} \quad (1.77)$$

Thus the impedance Z_2 is synthesized as shown in Fig. 1.28, in which the values of the components are in ohms and farads.

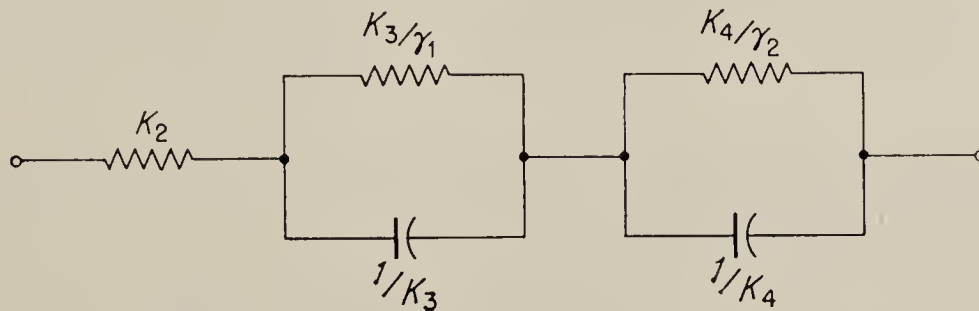


FIG. 1.28. Synthesis of Z_2 (first method).

The second method of synthesizing the impedance function is again illustrated by using Z_2 of Eq. (1.73) as an example. We write

¹ Guillemin, "Communication Networks," John Wiley & Sons, Inc., New York, 1948, vol. 2, pp. 211-216.

$$\begin{aligned}
 \frac{1}{Z_2} = Y_2 &= \frac{s}{K_2} \frac{(s + \gamma_1)(s + \gamma_2)}{s(s + \alpha_2)(s + \alpha_3)} \\
 &= \frac{s}{K_2} \left(\frac{A_1}{s} + \frac{A_2}{s + \alpha_2} + \frac{A_3}{s + \alpha_3} \right) \\
 &= B_1 + \frac{B_2 s}{s + \alpha_2} + \frac{B_3 s}{s + \alpha_3}
 \end{aligned} \tag{1.78}$$

where $B_1 = A_1/K_2$, etc. The form of Eq. (1.78) suggests that the admittance Y_2 consists of three components in parallel. The first of these is the conductance B_1 , and the other two both consist of a capacitor and resistor in series, since the admittance of such a combination is given by

$$Y = \frac{Cs/R}{Cs + (1/R)} \tag{1.79}$$

Hence by the second method the impedance is synthesized in the form shown in Fig. 1.29, with resistance and capacitance values again given in ohms and farads.

We may now complete the design of the transfer function given in Eq. (1.68). Using the first method of driving-point-impedance synthesis, we obtain the network of Fig. 1.30, in which the values of resistances and capacitances are given in ohms and farads, and where K_1/K_2 is given by Eq. (1.74).

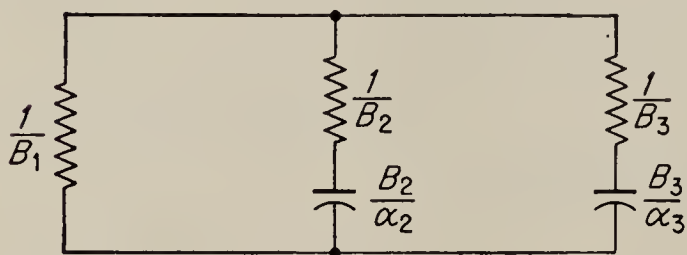


FIG. 1.29. Synthesis of Z_2 (second method).

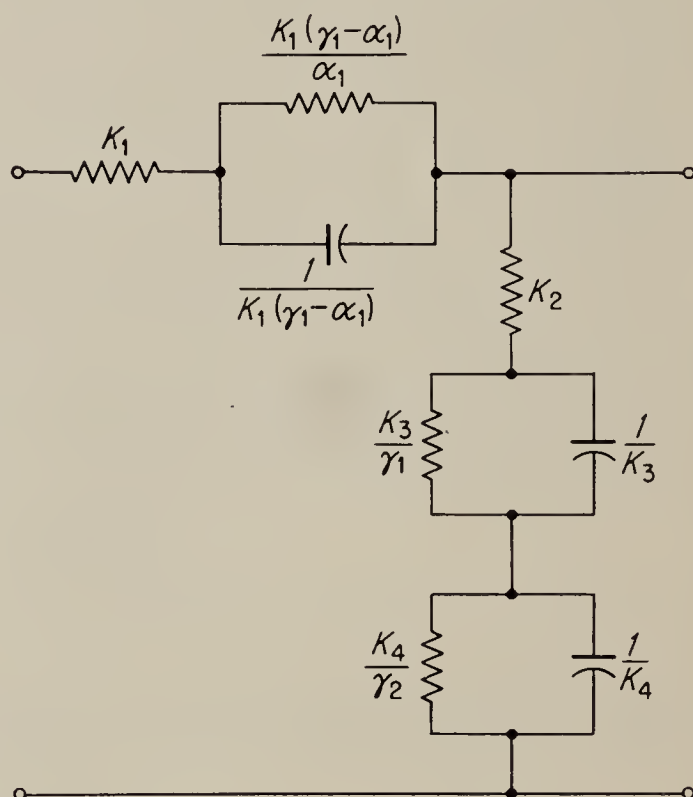


FIG. 1.30. Network having transfer function of Eq. (1.68).

1.25. Numerical Example. As an example of the synthesis method, consider the transfer function

$$H = h_0 \frac{(s + 10)(s + 20)(s + 50)^2}{(s + 1)(s + 2)(s + 200)^2(s + 500)} \tag{1.80}$$

In this function h_0 is a variable gain parameter which is to be maximized in the design. The asymptotic α diagram of this transfer function is shown in Fig. 1.31a. The function has four successive zeros and three successive poles, and can therefore be realized by a two-section ladder net-

work. In Fig. 1.31*b* and *c* are shown the α diagrams of two L sections, which together will realize the transfer function of Eq. (1.80). The two component transfer functions are

$$\begin{aligned} H_x &= h_x \frac{(s + 20)(s + 50)}{(s + 1)(s + 200)(s + 500)} \\ H_y &= h_y \frac{(s + 10)(s + 50)}{(s + 2)(s + 200)} \end{aligned} \quad (1.81)$$

In making the distribution of poles and zeros between the two component network sections, we keep in mind that the gain level between two successive poles has an upper bound for realizable transfer functions (see Fig.

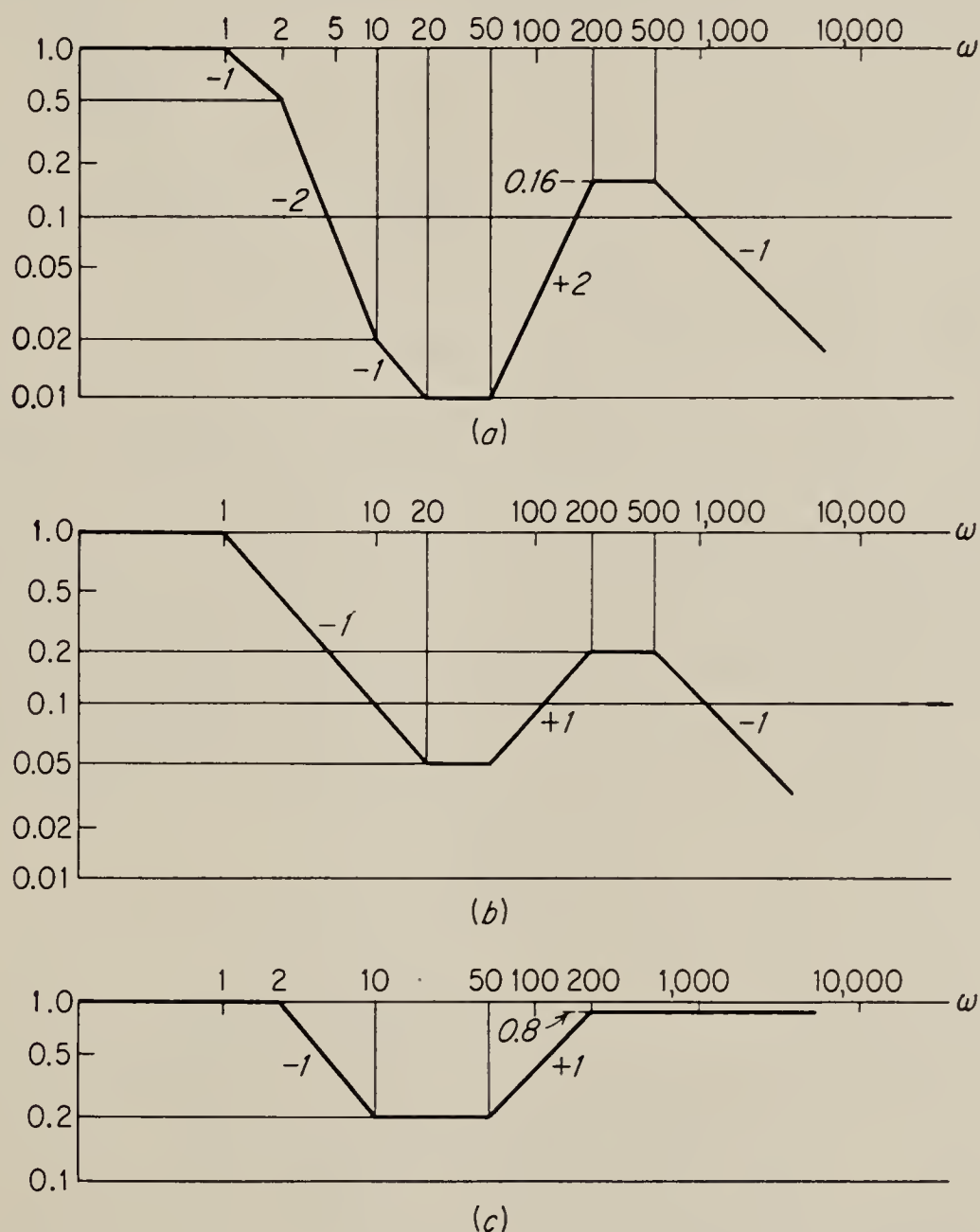


FIG. 1.31, α diagrams of over-all transfer function and component transfer functions.

1.24). The over-all network gain is therefore maximized by keeping the required gain level between the two poles (as between $s = -200$ and $s = -500$ in Fig. 1.31) as low as possible. For this reason the pole at $s = -1$ and the zero at $s = -20$ are assigned to H_1 rather than to H_2 . It should be noted that there is in general some optimum division of this sort, which occasionally may require the introduction into the component

networks of poles and zeros that do not exist in the over-all transfer function. The reader is referred to some of the problems at the end of this chapter which illustrate these principles (see, for instance, Prob. 1.14a).

We proceed by first synthesizing H_x . A sketch of $F(s)$ for the transfer function H_x is shown in Fig. 1.32, with the horizontal line for the maximum

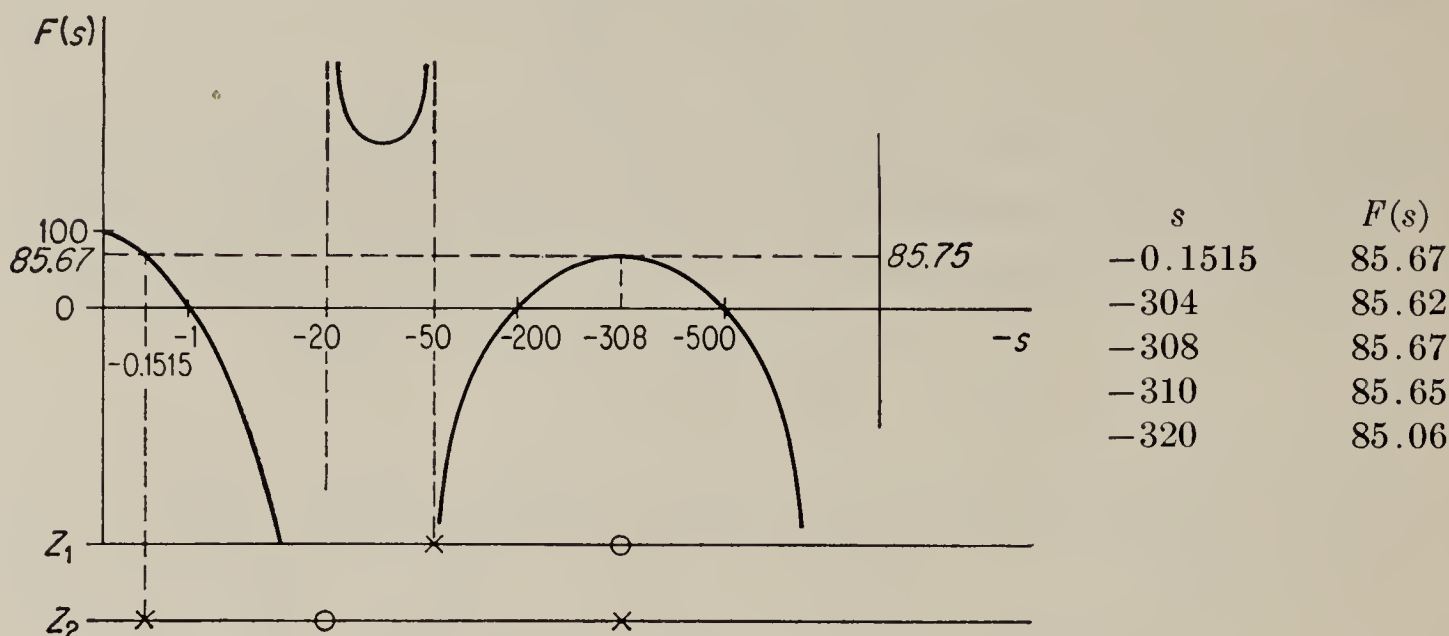


FIG. 1.32. Design of H_x .

value of h_x indicated. A few values of the trial-and-error solution for the peak value of $F(s)$ are shown in the table at the right. They give a maximum value of h_x of 85.67, with the maximum occurring at $s = -308$. Also the intersection of the line for $h_x = 85.67$ with $F(s)$ is found by trial and error to occur at $s = -0.1515$. Also indicated on Fig. 1.32 are the zeros and poles of Z_1 and Z_2 . The two poles of $F(s)$ at $s = -20$ and $s = -50$ could have been interchanged without significant difference in the final result, but otherwise the arrangement must be as shown. We have, therefore,

$$\begin{aligned} Z_{1x} &= K_{1x} \frac{s + 308}{s + 50} \\ Z_{2x} &= K_{2x} \frac{s + 20}{(s + 0.1515)(s + 308)} \end{aligned} \quad (1.82)$$

The ratio K_{1x}/K_{2x} is best evaluated for $s = 0$; from Eq. (1.71)

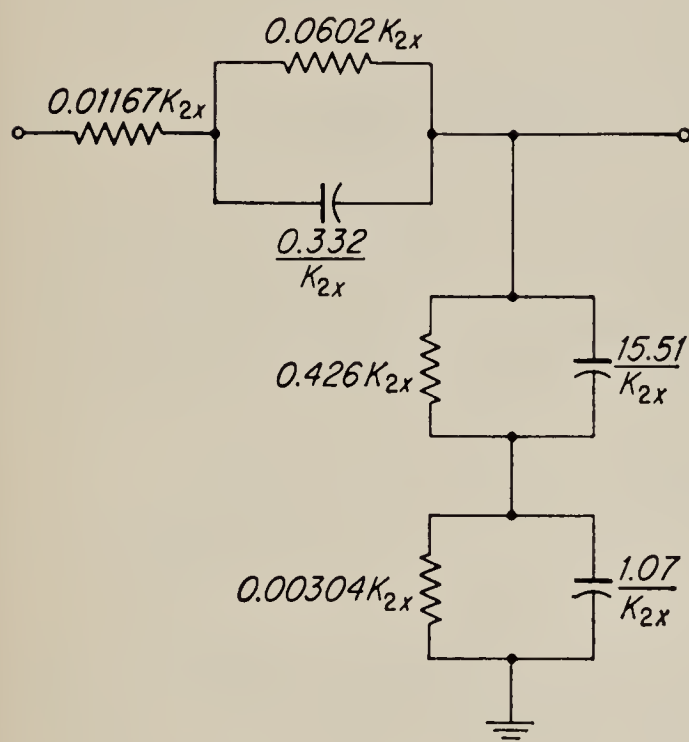
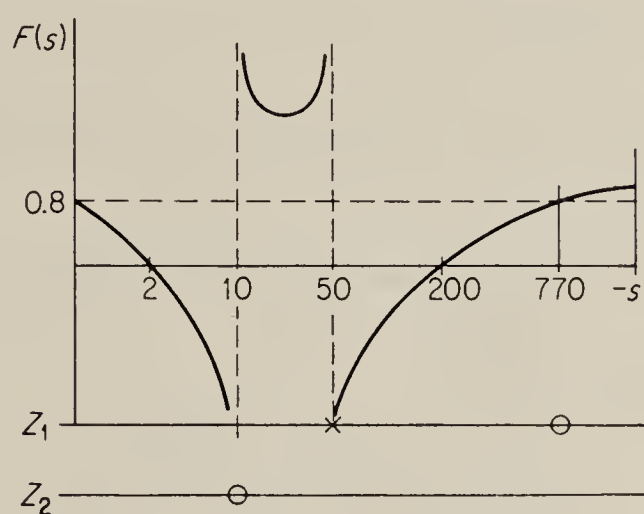
$$\begin{aligned} \frac{Z_{1x}}{Z_{2x}} \bigg|_{s=0} &= \frac{K_{1x}}{K_{2x}} \frac{(308)^2 \times 0.1515}{50 \times 20} = \frac{1}{85.67} (100 - 85.67) \\ \frac{K_{1x}}{K_{2x}} &= 0.01167 \end{aligned} \quad (1.83)$$

The partial-fraction expansion of the Z 's is

$$\begin{aligned} Z_{1x} &= 0.01167 K_{2x} \left(1 + \frac{258}{s + 50} \right) \\ Z_{2x} &= K_{2x} \left(\frac{0.06447}{s + 0.1515} + \frac{0.935}{s + 308} \right) \end{aligned}$$

The complete design of H_x is shown in Fig. 1.33, in which the values are expressed in ohms and farads. K_{2x} is left arbitrary to permit adjustment of the impedance level.

The design of H_y proceeds in a completely analogous manner. A sketch of $F(s)$ is shown in Fig. 1.34. The maximum value of h_{\max} in this case is $F(0)$, because larger values of h would require zeros or poles in one of the impedances for positive value of s . This is ruled out by the restrictions on realizable impedance functions. The intersection of the horizontal line for h_{\max} with $F(s)$ at $s = -770$ can be easily determined in this example by a direct algebraic solution. The distribution of zeros and poles


 FIG. 1.33. Final form of H_x .

 FIG. 1.34. Synthesis of H_y .

to the two impedances shown in the figure minimizes the number of capacitors. We have

$$\begin{aligned} Z_{1y} &= K_{1y} \frac{s + 770}{s + 50} \\ Z_{2y} &= K_{2y} \frac{s + 10}{s} \end{aligned} \quad (1.84)$$

The ratio K_{1y}/K_{2y} may be evaluated for $s = -\infty$ and is given by

$$\frac{K_{1y}}{K_{2y}} = \frac{1}{0.8} (1 - 0.8) = 0.25 \quad (1.85)$$

Hence

$$\begin{aligned} Z_{1y} &= 0.25K_{2y} \left(1 + \frac{720}{s + 50} \right) \\ Z_{2y} &= K_{2y} + \frac{10K_{2y}}{s} \end{aligned}$$

The final design of H_y is shown in Fig. 1.35, with resistances and capacitances in ohms and farads.

The final step in the design is to determine the ratio of K_{2x} to K_{2y} so that, when H_x and H_y are connected, the second L section will not load

the first one excessively (see Sec. 1.19). In general, this step calls for a calculation of the ratio of the impedances of the two sections, as required by Eq. (1.34). Since one order of connection may have advantages in the size of capacitors required, it is good practice to compute the circuit for both orders of connection.

The impedance ratio Z_{ox}/Z_{iy} for the case when network H_x precedes H_y is most easily found from Eq. (1.38).

$$\begin{aligned}\frac{Z_{ox}}{Z_{iy}} &= \frac{Z_{1x}H_xH_y}{Z_{2y}} = \frac{Z_{1x}}{Z_{2y}} H \\ &= \frac{0.7998K_{2x}s(s+20)(s+50)(s+308)}{K_{2y}(s+1)(s+2)(s+200)^2(s+500)}\end{aligned}\quad (1.86)$$

An asymptotic plot of this function shows the maximum to occur for $1 < \omega < 2$ with a value of about $0.006K_{2x}/K_{2y}$. If the impedance ratio is to be less than 0.1 for all frequencies, the ratio of K_{2y} to K_{2x} should be greater than 0.06. The network resulting for a ratio of 0.06 is shown in Fig. 1.36.

If H_y precedes H_x , we consider the ratio

$$\begin{aligned}\frac{Z_{oy}}{Z_{ix}} &= \frac{Z_{1y}}{Z_{2x}} H \\ &= 17.13 \frac{K_{2y}}{K_{2x}} \frac{(s+0.1515)(s+10)(s+50)(s+308)(s+770)}{(s+1)(s+2)(s+200)^2(s+500)}\end{aligned}\quad (1.87)$$

The maximum value again occurs for $1 < \omega < 2$ and has a value of less than $40K_{2y}/K_{2x}$. Thus, if the ratio of the impedances Z_{oy}/Z_{ix} is to be less than 0.1, $K_{2x}/K_{2y} \geq 400$. The resulting network is shown in Fig. 1.37.

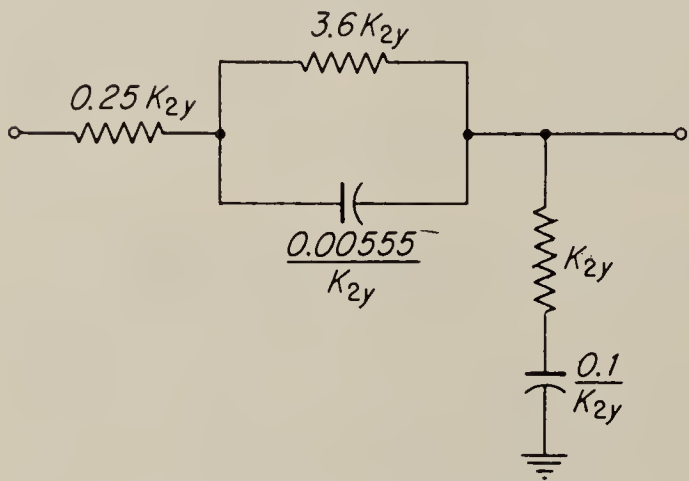
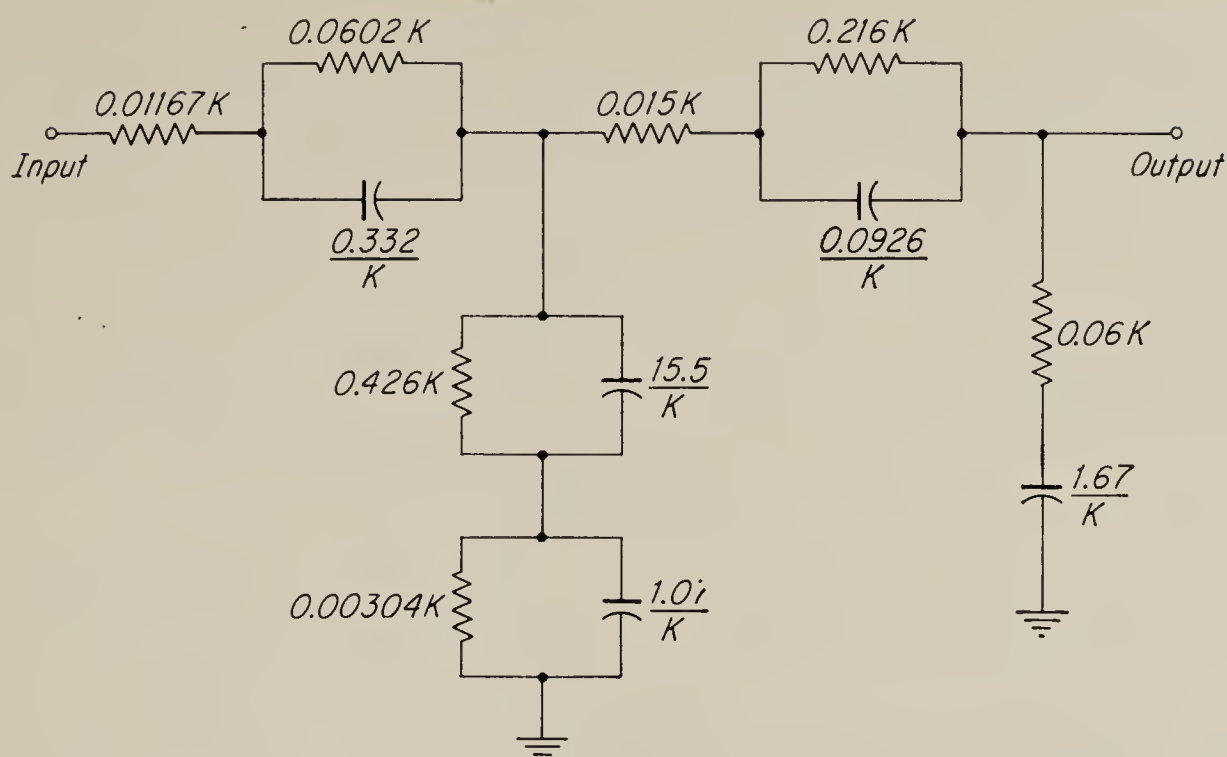
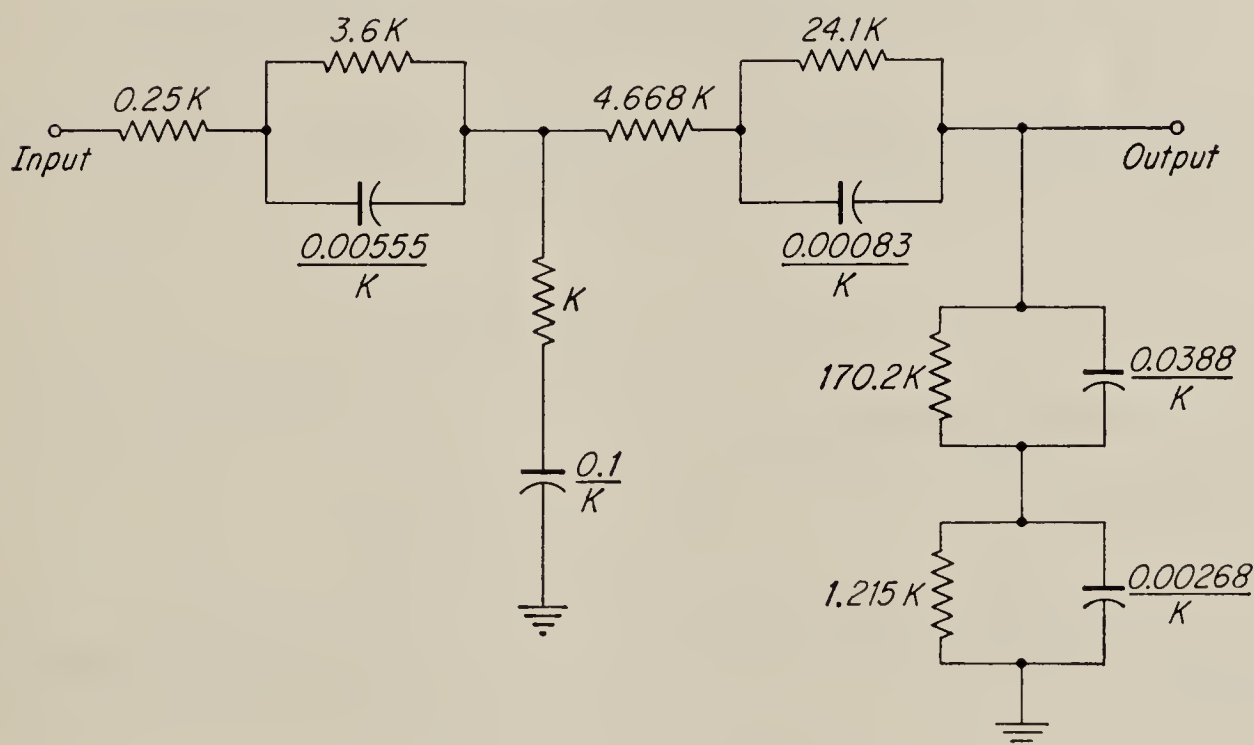


FIG. 1.35. Final form of H_y .

The final decision as to which circuit is better depends now on practical circuit considerations. Thus, if the network were to drive the grid of a vacuum tube, a maximum d-c resistance between grid and ground of 1 megohm might be specified. This specification would be met in the circuit of Fig. 1.36 with

a K of about 3×10^6 , while in Fig. 1.37 K would be about 3×10^4 . (It is assumed that the network driving source has zero d-c resistance.) Under these conditions the largest capacitance in Fig. 1.36 would be about $5.1 \mu\text{f}$, while in Fig. 1.37 it would be only $3.3 \mu\text{f}$. Since large capacitors are bulky, the network of Fig. 1.37 has a slight advantage. In general, it will be found that the size of the largest capacitors is minimized by

FIG. 1.36. Final network with H_x preceding H_y .FIG. 1.37. Final network with H_y preceding H_x .

placing the network section with the lowest frequency poles toward the end of the cascade.

1.26. Bridged-T and Twin-T Networks.¹ In this section we discuss two networks that are commonly used to obtain transfer functions with

¹ Chestnut and Mayer, "Servomechanisms and Regulating System Design," John Wiley & Sons, Inc., New York, 1955, vol. II, pp. 187–209. Valley and Wallman, "Vacuum Tube Amplifiers," Radiation Laboratory Series, vol. 18, McGraw-Hill Book Company, Inc., New York, 1948, pp. 384–391. James, Nichols, and Phillips, "Theory of Servomechanisms," Radiation Laboratory Series, vol. 25, McGraw-Hill Book Company, Inc., New York, 1947, pp. 117–124. L. Stanton, Theory and Application of Parallel-T Resistance Capacitance Frequency Selective Networks, *Proc. IRE*, vol. 34, pp. 447–456, 1946.

complex zeros. The bridged T is the simpler of the two networks, but the complex zeros produced by it can have only a relatively small imaginary part. The twin T is more complicated but also more flexible, permitting the synthesis of transfer functions with purely imaginary zeros

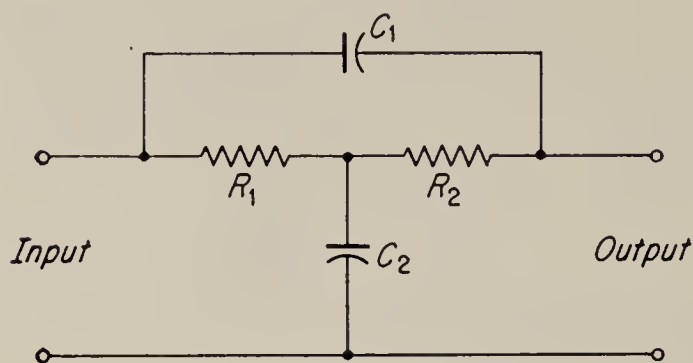


FIG. 1.38. Bridged-T network.

(i.e., complete rejection of one frequency). The twin T also can be designed with a nonminimum phase type of transfer function; i.e., it may have zeros with positive real parts.

The form of the bridged T is shown in Fig. 1.38. The transfer function can be found by applica-

tion of standard techniques (the node method gives the result most rapidly) and can be put into the form

$$\frac{\hat{e}_o}{\hat{e}_i} = H(s) = \frac{R_1 R_2 C_1 C_2 s^2 + C_1 (R_1 + R_2) s + 1}{R_1 R_2 C_1 C_2 s^2 + [C_1 (R_1 + R_2) + R_1 C_2] s + 1} \quad (1.88)$$

By differentiation of the magnitude of the frequency function $|H(j\omega)|$, it is found that the function has a minimum for

$$\omega = \omega_c = \frac{1}{\sqrt{R_1 R_2 C_1 C_2}} \quad (1.89)$$

It is convenient, therefore, to normalize the frequency variable s by defining

$$p = \frac{s}{\omega_c} = s \sqrt{R_1 R_2 C_1 C_2} \quad (1.90)$$

The transfer function is simplified further by making use of the additional definitions:

$$C_1 (R_1 + R_2) = \frac{2\zeta}{\omega_c}$$

and

$$\frac{R_1}{R_2} = \alpha \quad (1.91)$$

With these definitions the transfer function may then be put in the form

$$H(p) = \frac{p^2 + 2\zeta p + 1}{p^2 + \left(2\zeta + \frac{\alpha + 1}{2\zeta}\right) p + 1} \quad (1.92)$$

Physical realizability requires both ζ and α to be positive real numbers. Thus ζ cannot be made equal to zero, and the network cannot completely reject the frequency ω_c . In fact, it is seen from Eq. (1.92) that, as ζ and therefore the minimum gain of the network are decreased, the two poles

of the transfer function move farther and farther apart, and therefore the phase lag introduced by the network at frequencies well below ω_c becomes larger. This is illustrated by the attenuation and phase-shift curves

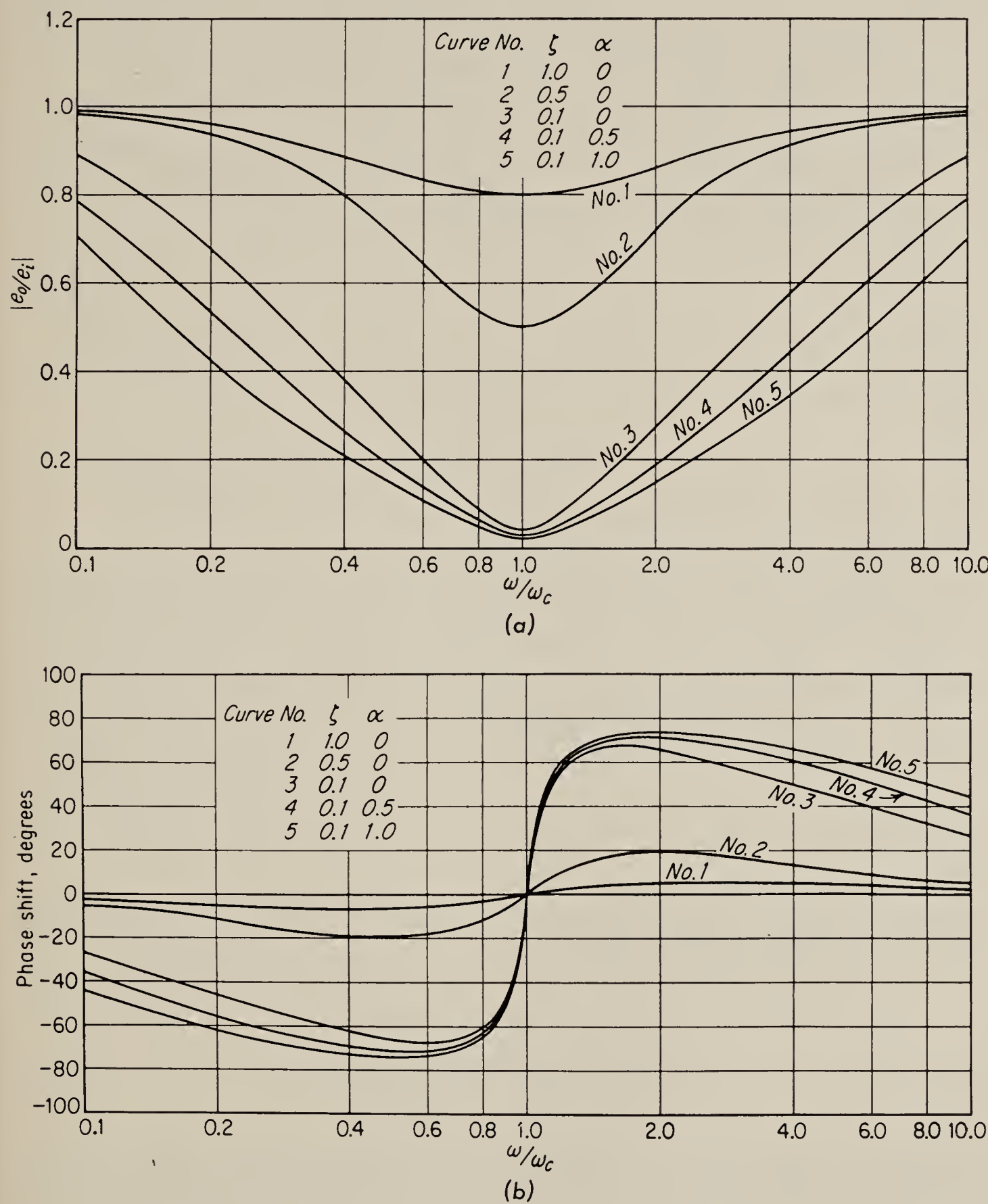


FIG. 1.39. Transfer characteristics of bridged-T network.

given in Fig. 1.39. Since the complex zeros produced by the network are quite often used to cancel a set of complex poles produced by some other component of the loop, the large phase lag produced at low frequencies is sometimes objectionable.

It should be mentioned that, although the bridged-T network of the

form shown in Fig. 1.38 seems to be most commonly used, it is also possible to construct the network with resistors and capacitors interchanged.

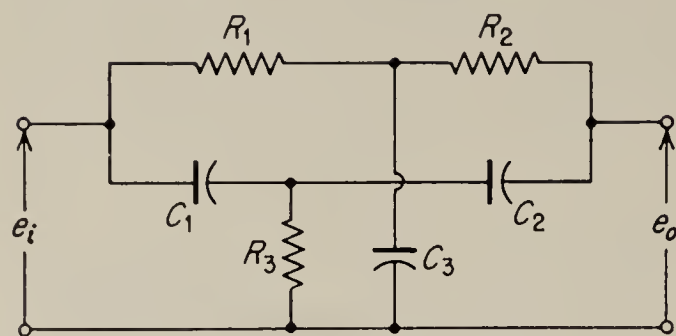


FIG. 1.40. Twin-T network.

The basic form of the transfer function as given by Eq. (1.92) is not changed by this procedure.

The twin-T network is shown in Fig. 1.40. Its transfer function is again most easily found by use of the node method and can be put into the form

$$\frac{\hat{e}_o}{\hat{e}_i} = H(s) = \frac{N(s)}{\Delta(s)} \quad (1.93)$$

where

$$N(s) = R_1 R_2 R_3 C_1 C_2 C_3 s^3 + (R_1 + R_2) R_3 C_1 C_2 s^2 + R_3 (C_1 + C_2) s + 1 \quad (1.94)$$

and

$$\Delta(s) = R_1 R_2 R_3 C_1 C_2 C_3 s^3 + [(R_1 + R_2) R_3 C_1 C_2 + (R_2 + R_3) R_1 C_2 C_3 + R_1 R_3 C_1 C_3] s^2 + [R_1 (C_2 + C_3) + R_2 C_2 + R_3 (C_1 + C_2)] s + 1 \quad (1.95)$$

Since the twin T is most often used as a rejection network, it is desirable to determine, first, under what conditions there is a null in the transfer function. The numerator of the transfer function is a cubic polynomial and therefore has three zeros. If the network is to have a null at the frequency ω_c , two of the zeros must be $j\omega_c$ and $-j\omega_c$, and $N(s)$ must have the form

$$N(s) = \left(\frac{s}{j\omega_c} + 1 \right) \left(\frac{s}{-j\omega_c} + 1 \right) \left(\frac{s}{\gamma} + 1 \right) = \frac{s^3}{\omega_c^2 \gamma} + \frac{s^2}{\omega_c^2} + \frac{s}{\gamma} + 1 \quad (1.96)$$

where γ is the third zero, which must be negative and real. Comparison with Eq. (1.94) indicates that

$$\frac{1}{\omega_c^2} = (R_1 + R_2) R_3 C_1 C_2 \quad (1.97)$$

$$\frac{1}{\omega_c^2} = \frac{R_1 R_2 R_3 C_1 C_2 C_3}{R_3 (C_1 + C_2)} \quad (1.98)$$

The second relation is obtained by dividing the coefficient of s^3 by that of s . Equating Eqs. (1.97) and (1.98) and canceling common factors gives for the null condition

$$\frac{C_3}{C_1 + C_2} = \frac{R_3 (R_1 + R_2)}{R_1 R_2} = n \quad (1.99)$$

where n is any real, positive number. It will be shown presently that n has an optimum value depending on the ratio of R_1 to R_2 and that of C_1 to

C_2 . If the network is symmetrical, i.e., if $R_1 = R_2$ and $C_1 = C_2$, it will be found that the optimum value for n is 1.

Note that, if the three capacitors are fixed, the adjustment of any one of the three resistors will suffice to meet the null condition and cause the circuit to reject one frequency. However, if the null is to occur at a prescribed frequency, at least two resistors must be adjusted simultaneously, and the variation of one of the resistors is a fairly complicated function of the value of the other. In practice the circuit is quite often constructed using a three-gang potentiometer for the three resistances. The three resistances are then all equal, so that the null condition is met with $n = 2$. Variation of the three resistors permits a smooth variation of the rejection frequency.

If the null condition is met, the transfer function of the circuit can be simplified considerably. To do this, we find it convenient to let

$$C_2 = \alpha C_1, \quad R_2 = \beta R_1 \quad (1.100)$$

Substituting for R_2 in Eq. (1.99), we obtain

$$R_3 = \frac{n\beta}{1 + \beta} R_1 \quad (1.101)$$

and using this result and Eq. (1.100) in Eq. (1.97), it is found that

$$R_1 C_1 = \frac{1}{\omega_c} \sqrt{\frac{1}{\alpha\beta n}} \quad (1.102)$$

Finally, the frequency variable s is again normalized with respect to the rejection frequency ω_c by defining

$$p = \frac{s}{\omega_c} \quad (1.103)$$

Substitution of all of these definitions and results [Eqs. (1.100) to (1.103)] into Eq. (1.93) yields then, after some algebraic manipulation, the transfer function

$$H\left(\frac{s}{\omega_c}\right) = H(p) = \frac{1 + p^2}{1 + Kp + p^2} \quad (1.104)$$

$$\text{where} \quad K = \frac{n(1 + \alpha) + \alpha(1 + \beta)}{\sqrt{\alpha\beta n}} \quad (1.105)$$

Thus, for a given null frequency the circuit behavior is completely determined by n , by the ratio of R_2 to R_1 , and by the ratio of C_2 to C_1 .

Normally it is desirable to have as sharp a null as possible at the rejection frequency. The sharpness of the null may be defined in a number of ways, but a convenient one is to consider the slope of the magnitude ratio $|H(j\omega/\omega_c)|$ at the null as a measure of sharpness. Accordingly we

define the optimum value of K as the value giving the steepest slope of $|H(j\omega/\omega_c)|$ for $\omega = \omega_c$. By differentiation of the magnitude of Eq. (1.104) with p set equal to $j\omega/\omega_c$, it is found that the slope at $\omega/\omega_c = 1$ is $-2/K$. Hence the optimum value of K is the smallest realizable value. In determining this value, we must keep in mind that α , β , and n must be positive real numbers. It is convenient to express K as the sum of two terms:

$$K = \left(\sqrt{\frac{n}{\alpha\beta}} + \sqrt{\frac{\alpha\beta}{n}} \right) + \sqrt{\frac{\alpha}{\beta}} \left(\sqrt{n} + \frac{1}{\sqrt{n}} \right) \quad (1.106)$$

Note that the two terms indicated by the parentheses are both positive and that K is therefore minimized by minimizing each one separately. The first term has the form $(x + 1/x)$. By differentiation this function is easily shown to be minimum for $x = 1$. Hence one condition for minimum K is

$$\frac{\alpha\beta}{n} = 1 \quad (1.107)$$

or

$$n = \alpha\beta$$

Substituting this value of n into the expression for K gives

$$K = 2 + \alpha + \frac{1}{\beta} \quad (1.108)$$

It is clear that the minimum value of K is 2 and that it is approached by making α as small as possible and β as large as possible. In practice, if the twin T drives a grid of a vacuum tube, R_2 should normally not exceed 1 megohm. Also, since R_1 should be about ten times as large as the source impedance of the input, a practical minimum on R_1 is about 10,000 ohms. Hence a practical maximum value for β is about 100. Similar considerations limit α to about 0.01. Hence the practical minimum for K is of the order of 2.02.

The twin T is often built symmetrically, i.e., with $\alpha = \beta = 1$. Under these conditions the best value of n is seen to be 1, and $K = 4$, or twice its optimum value. As can be seen in Fig. 1.41, this value of K produces a decidedly larger phase lag at frequencies below the rejection frequency than the optimum value does.

Both the bridged T and the twin T are used to some extent as series equalizers in a-c servo systems. When used in this way, the bridged-T network has an effect similar to that of a standard d-c lead network (see Sec. 1.13), while the twin T produces pure differentiation; i.e., the output signal amplitude is proportional to signal frequency, and the output phase leads the input phase by 90° .

The reason for this may be explained as follows: Consider first a typical all-a-c system, such as the one shown in Fig. 7.1. If the shaft of

the input synchro in that figure is given a displacement varying sinusoidally with time, then, as is shown in Sec. 5.2, the error signal going into the amplifier has the form

$$e = E_m (\cos \omega_s t \cos \omega_c t)$$

(1.109)

where ω_s is the relatively low signal frequency at which the shaft of the synchro is being oscillated, while ω_c is the carrier frequency, which is

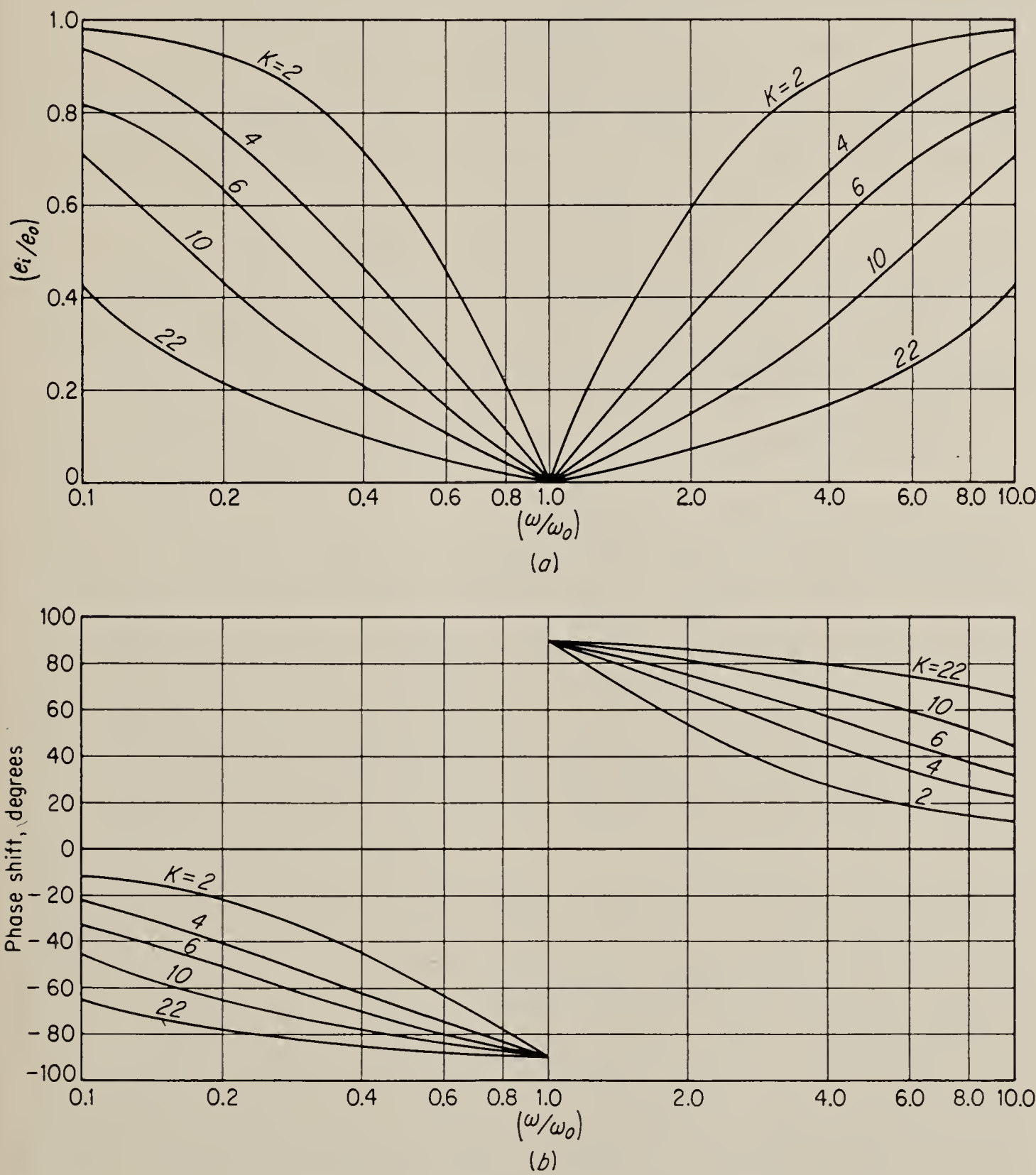


FIG. 1.41. Characteristics of twin T for various values of K.

assumed to be at least an order of magnitude greater than the signal frequency. Furthermore, it is shown in Sec. 7.4 that to a first approximation the torque generated by a two-phase servomotor supplied with a sig-

nal of the form of Eq. (1.109) is given by

$$T = T_m \cos \omega_s t \quad (1.110)$$

i.e., it is proportional only to the signal envelope. An equalizing network inserted between the synchro and the motor to provide phase lead at signal frequencies should, therefore, advance the phase of the signal envelope. Hence, if the input to the network is a signal of the form given by Eq. (1.109), the output should be

$$e_o = E_{om} \cos (\omega_s t + \theta) \cos \omega_c t \quad (1.111)$$

where θ is the desired phase lead. By use of standard trigonometric identities the input signal [Eq. (1.109)] may be written in the form

$$e = \frac{E_{om}}{2} [\cos (\omega_c - \omega_s)t + \cos (\omega_c + \omega_s)t] \quad (1.112)$$

showing that the signal consists of two sidebands, one at the frequency $\omega_c - \omega_s$ and the other at the frequency $\omega_c + \omega_s$. Similarly the output signal may be expanded into the form

$$e_o = \frac{E_{om}}{2} \{ \cos [(\omega_c - \omega_s)t - \theta] + \cos [(\omega_c + \omega_s)t + \theta] \} \quad (1.113)$$

Thus we see that the function of the network must be to provide phase lag at frequencies below the carrier frequency and phase lead at frequencies above the carrier frequency. Inspection of Figs. 1.39 and 1.41 indicates that both the bridged T and the twin T will perform this function, provided that they are tuned so that their null frequency coincides exactly with the carrier frequency. Note, however, that they perform the operation only approximately, since their characteristics are symmetrical with respect to the logarithm of frequency, while the desired characteristic, as shown by Eq. (1.113), should be symmetrical with respect to the frequency itself. This is not, however, a serious objection if the signal frequencies are low with respect to the carrier frequency. A much more serious difficulty with this type of series equalization is that the performance depends very critically on the carrier frequency, and even relatively small shifts from the nominal value result in a serious loss of performance. This is due to the fact that the network characteristics depend on the absolute frequency of the signal rather than on its ratio to the carrier frequency.

To examine this problem quantitatively, and also to indicate an approach toward a design procedure, suppose that the signal frequency is given by

$$\omega_s = c\omega_c \quad (1.114)$$

Then the upper- and lower-sideband frequencies are, respectively, $\omega_c(1 + c)$ and $\omega_c(1 - c)$. The discussion is illustrated by using the

bridged-T network but is easily extended to the twin T (see Prob. 1.19). The transfer function is given in Eq. (1.92). In order to get the frequency function, let $p = ju = j\omega/\omega_c$, so that the normalized upper- and lower-sideband frequencies become simply $j(1 + c)$ and $j(1 - c)$, respectively. Substituting into Eq. (1.92) and replacing $2\zeta + (\alpha + 1)/2\zeta$ simply by $2\zeta_d$, we obtain, after some algebraic rearranging,

$$H(ju) = \frac{\zeta}{\zeta_d} \frac{1 + j \frac{c(c + 2)}{2\zeta(c + 1)}}{1 + j \frac{c(c + 2)}{2\zeta_d(c + 1)}} \quad (1.115)$$

for the upper sideband and

$$H(ju) = \frac{\zeta}{\zeta_d} \frac{1 - j \frac{c(c - 2)}{2\zeta(c - 1)}}{1 - j \frac{c(c - 2)}{2\zeta_d(c - 1)}} \quad (1.116)$$

for the lower sideband. When the network is to be used to provide phase lead in a servo, ζ_d is made very much larger than ζ . Hence, for small c the denominator is approximately equal to unity, and the transfer function becomes approximately

$$H(j\omega) \approx \frac{\zeta}{\zeta_d} \left[1 \pm j \frac{c(c \pm 2)}{2\zeta(c \pm 1)} \right] \approx \frac{\zeta}{\zeta_d} \left(1 \pm j \frac{c}{\zeta} \right) \quad (1.117)$$

An approximate design procedure now becomes apparent. Suppose, for example, that the carrier frequency of the servo is 60 cps and that it is desired to provide a phase lead of $45^\circ = \pi/4$ radian at a signal frequency of 3 cps. Then $c = 0.05$, and since for a 45° angle the real and imaginary parts are equal, $\zeta = 0.05$. From Eq. (1.92) we have that, for the network to be physically realizable, ζ_d must be greater than 20. This fulfills the requirement above that $\zeta_d \gg \zeta$. Note, however, that the gain of the network at the carrier frequency ($c = 0$) is ζ/ζ_d , which in this case is less than $1/400$. Thus, in order to provide a reasonable phase lead at low signal frequencies, the network must introduce a large amount of attenuation. Furthermore it is clear that the equalizer characteristics available from this type of network cannot compare in variety or complexity with the characteristics of d-c equalizer networks discussed earlier in this chapter. A-c servos therefore tend to have poorer performance than servos in which equalization is applied to the signal frequencies directly.

The effect of carrier-frequency shift on network performance can now also be illustrated by an example. Suppose that the frequency shifts 5 per cent, or 3 cps. Then as far as the network is concerned, c for the upper sideband becomes $6/60 = 0.1$, while for the lower sideband c is zero.

With the same value of ζ as before, the phase lead of the upper sideband is about 63.5° , while the phase of the lower sideband is zero. By inverting the process used to derive Eq. (1.113), it can be shown that the signal-frequency phase shift is approximately equal to one-half the difference between the phase shifts of the sidebands, which in this case amounts to about 31° . Thus a small shift in carrier frequency results in a large loss of phase lead. A further effect is that the carrier phase is also shifted by about 30° . As shown in Chap. 7, this results in reduction of torque generated by the motor and excessive heating.

Although the commercial 60-cps supply frequency is usually maintained very close to its nominal value, this is not true of most 400-cps-circuit supplies. The frequency of currently available aircraft supplies cannot be guaranteed to be regulated to better than ± 10 per cent, or ± 40 cps. Such a frequency variation will make networks of the type discussed here quite useless. The regulation problem of aircraft supplies may be solved in the next few years, thus making high-performance a-c servos practical in aircraft.

1.27. The Use of Amplifiers with RC Networks.¹ The variety of transfer functions available with RC networks can be extended greatly by using networks in conjunction with amplifiers. The subject is too large to be dealt with completely here, and only some of the possibilities are indicated by the following examples.

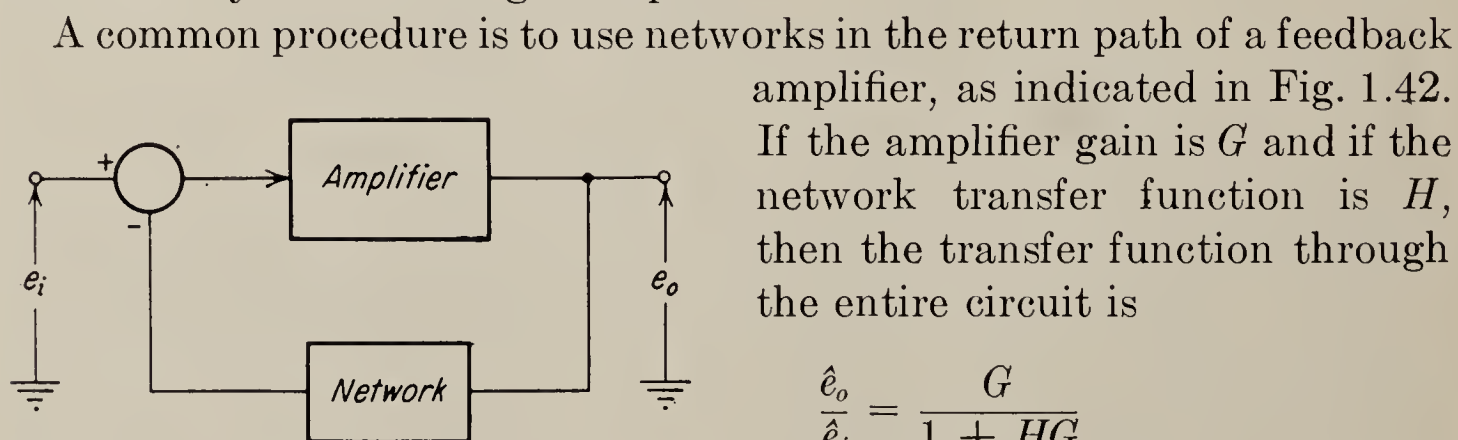


FIG. 1.42. Feedback amplifier with RC network in feedback path.

A common procedure is to use networks in the return path of a feedback amplifier, as indicated in Fig. 1.42. If the amplifier gain is G and if the network transfer function is H , then the transfer function through the entire circuit is

$$\begin{aligned} \frac{\hat{e}_o}{\hat{e}_i} &= \frac{G}{1 + HG} \\ &\approx \frac{1}{H} \quad \text{if } HG \gg 1 \end{aligned} \quad (1.118)$$

Thus, provided the loop gain is sufficiently high, the approximate transfer function is the reciprocal of the network transfer function.

This principle can be applied in the generation of transfer functions having complex poles by making the network in the return path of the feedback amplifier a twin T. Qualitatively such an amplifier will have a gain of unity at very low and very high frequencies, but at the null frequency of the twin T, the gain is G , since the feedback is not operative there. Clearly the transfer function is not the exact reciprocal of the twin-T transfer function. Even though the transfer function of the twin T

¹ Valley and Wallman, *op. cit.*, pp. 391–408.

has imaginary zeros, that of the feedback amplifier cannot have imaginary poles, since this would require infinite amplification at the null frequency. The discrepancy can, however, be made as small as desired by making the gain of the amplifier, G , large enough.

Another interesting application of the twin-T circuit in connection with a feedback amplifier is shown in Fig. 1.43. This circuit is referred to as the *rejection amplifier*, and it has the effect of sharpening the null of the

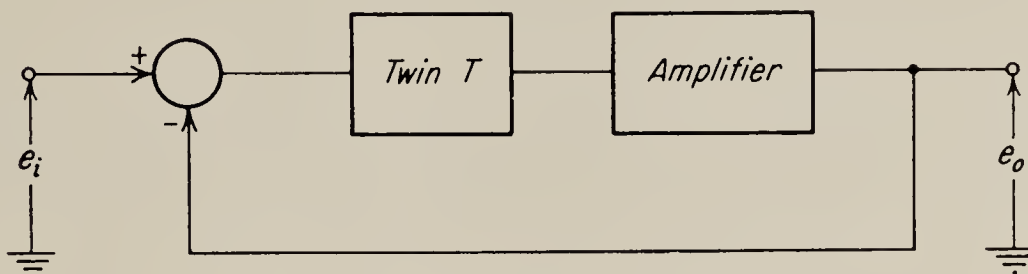


FIG. 1.43. Rejection amplifier.

twin T. The operation of this circuit is most easily explained by determining the over-all transfer function \hat{e}_o/\hat{e}_i . Application of standard techniques and the transfer function for the twin T [Eq. (1.104)] gives

$$\frac{\hat{e}_o}{\hat{e}_i} = \frac{G \frac{(p^2 + 1)}{(p^2 + Kp + 1)}}{1 + \frac{G(p^2 + 1)}{(p^2 + Kp + 1)}} = \left(\frac{G}{G + 1} \right) \left(\frac{p^2 + 1}{p^2 + \frac{K}{(G + 1)}p + 1} \right) \quad (1.119)$$

where G is the amplifier gain. Comparison with Eq. (1.104) indicates that the effect of the circuit is essentially to change the K in the denominator of the transfer function of the twin T to $K/(1 + G)$. In the discussion of the previous section it was shown that the slope of the magnitude of the frequency function at the null was $-2/K$ for the twin T; hence it is clear that the action of the amplifier is to make the slope steeper, and therefore to sharpen the null.

In both the examples given, the transfer functions that were obtained are of the type that could also be obtained by the use of inductances. Thus, complex poles are commonly obtained by simple RLC resonant circuits, while the sharp rejection characteristic of the second example may be shown to be identical to that of a bridged-T network in which one of the resistances is replaced by an inductance.¹ This bears out the statement made in Sec. 1.14 that, by the use of RC networks with amplifiers, practically any transfer function that can be realized with resistances, capacitances, and inductances can be realized without actually having to use inductances.

A final example illustrates the use of a difference amplifier (see Sec. 2.11) to generate a nonminimum phase transfer function. The circuit is shown in Fig. 1.44.

¹ Valley and Wallman, *op. cit.*, p. 385.

The difference amplifier is supposed to have infinite input impedance, and its output voltage e_o is given by $G(e_1 - e_2)$. Then, since

$$\frac{\hat{e}_1}{\hat{e}_i} = \frac{1}{Ts + 1} \quad \text{and} \quad \frac{\hat{e}_2}{\hat{e}_i} = \frac{Ts}{Ts + 1}$$

(see Sec. 1.13), the over-all transfer function is

$$\frac{\hat{e}_o}{\hat{e}_i} = G \frac{1 - Ts}{1 + Ts} \quad (1.120)$$

This transfer function is of the *all-pass* type; i.e., its frequency ratio has a constant magnitude of G , but the phase shift between output and input changes from 0 to 180°. This transfer function can also be obtained,

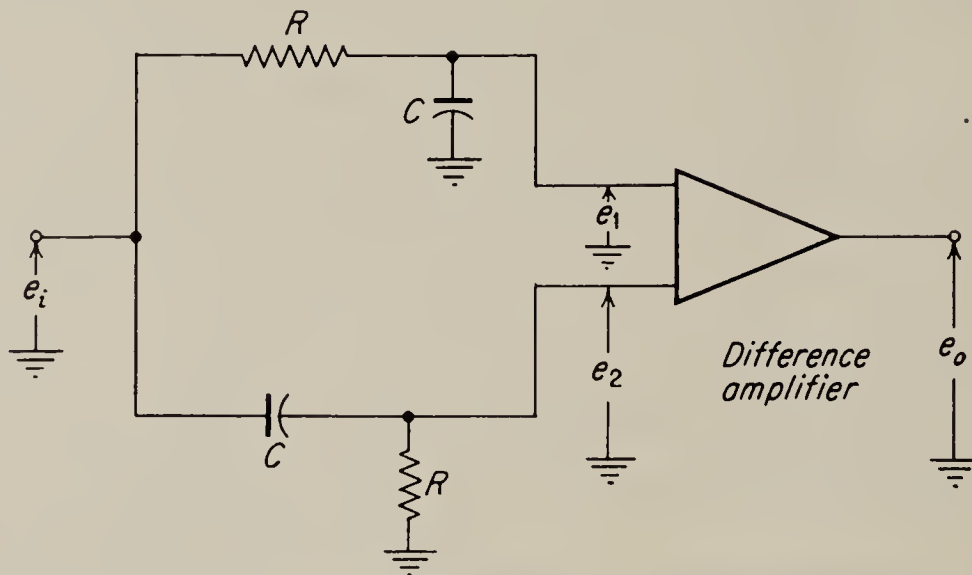


FIG. 1.44. Use of difference amplifier to produce nonminimum phase transfer function.

perhaps more simply, by a bridge network in which each arm contains either a resistor R or a capacitor C , and in which the R 's and C 's alternate. The disadvantage of a bridge circuit is, however, that there is no common ground connection between output and input. Hence a circuit of the sort shown in Fig. 1.44 might be preferable.

1.28. Reliability. The great complexity of modern electronic equipment and, to a lesser extent, other forms of equipment and the resulting failures in operation have led in the past few years to an intensive study of the basic problem of reliability. This interest is especially great in the military services and was probably brought to a head when it was shown during the Korean War that at any one time, two-thirds of all the electronic equipment in the theatre was inoperative because of failure.

Indicative of this rapidly growing interest in reliability is the fact that in 1954 almost no contracts for military equipment explicitly mentioned reliability, while at the present time almost all contracts require proof of reliability by performance. Secondly, reliability is at the present time placed *first* in order of importance in design contracts by the military, even ahead of the meeting of performance specifications.

The major reasons for failure of electronic and other devices are four-

fold. First, complexity. Second, improper design and over-loading of components. Third, failure due to environmental conditions such as heat, humidity, and shock. Fourth, improper maintenance.

That equipment designed in the past few years is more complex than its predecessors is indicated by the fact that the number of vacuum tubes carried in military vehicles from aircraft carriers to intercontinental bombers has increased by a factor of 4 or 5 since World War II. Even if each element were just as or more reliable, this increase in complexity would cause a dramatic increase in the failure rate.

We define reliability as the probability that a device will operate properly under given environmental conditions for a specified length of time. The time specified may be 20 years for tubes installed in the transatlantic telephone cable, 100 hours for a long-range bomber, or 1 hour for a guided missile.

All of the elements in a piece of electronic equipment do not have the same reliability, of course. Resistors are perhaps 0.9999 reliable, while vacuum tubes are perhaps 0.7 or 0.8 reliable. It is inevitable that the vacuum tube will continue to be the most unreliable component in electronic equipment because of the high temperatures involved in its operation and the delicate nature of its construction. However, by proper design and quality control its reliability will certainly be improved considerably. To find the over-all reliability of the equipment, we need only take the product of the reliability of each component, provided we make the more or less realistic assumption that the reliability of one component does not affect the reliability of the other components. Since we are interested in only the first failure, this seems justified. If it is further assumed that all components are equally reliable, simply for demonstration's sake, the over-all reliability is the reliability of the individual component raised to a power equal to the number of components in the unit. In Fig. 1.45 is shown a plot of such a calculation. The required reliability of the individual components is plotted as a function of the desired over-all reliability of the equipment. The number of components in the equipment is the parameter. Notice, for example, that in a unit with 1,000 components, 99.9 per cent reliability of the individual elements will result in about 37 per cent reliability of the over-all unit, and if the individual reliability is raised to 99.99 per cent, the unit is still only 91 per cent reliable.

It is obvious from this chart that the designer must strive for simplicity and that a modification which requires a considerable increase in complexity should probably not be accepted even if its performance is considerably better than the simpler model's. It is apparent also that the list of preferred tube types must be narrowed considerably and then adhered to religiously if we are to achieve reliable electronic components.

Designers must be encouraged *not* to design new circuits for each new device but to use standard circuits that are mass-produced and under rigid quality control.

The second cause of unreliable devices is the attempt by the designer to wring every last bit of performance from electronic components. The statistical life history of a batch of components or complete units, whatever their type, usually falls into a pattern such as that shown in Fig. 1.46.

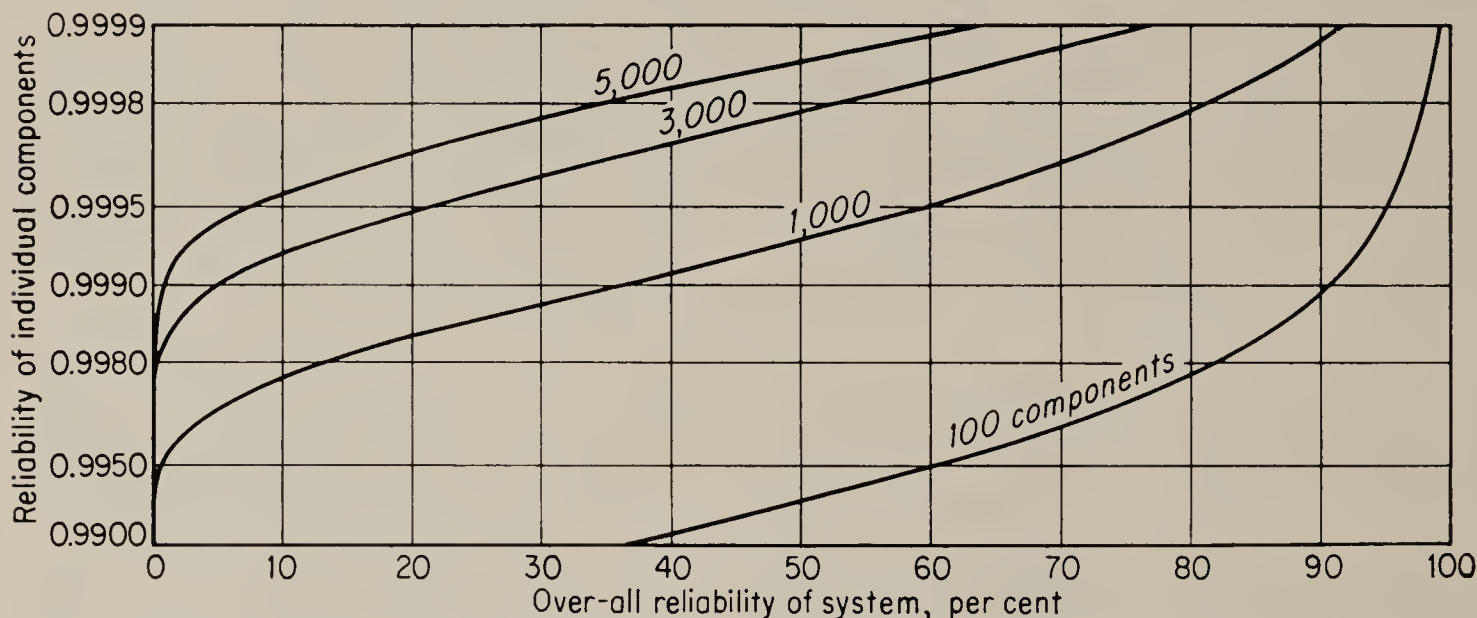


FIG. 1.45. Over-all reliability of a device as a function of the reliability of the individual components.

The curve may be divided into three regions. Region *A*, or the *infant mortality* region, consists of failures that occur in the break-in or shakedown period. The Navy's insistence on a shakedown cruise and specifications by buyers for 10 to 20 hours operation of a unit at the factory are both good procedures, as is seen from this chart. Next comes region *B*, the operating region. If the unit is well designed, the failures that occur in

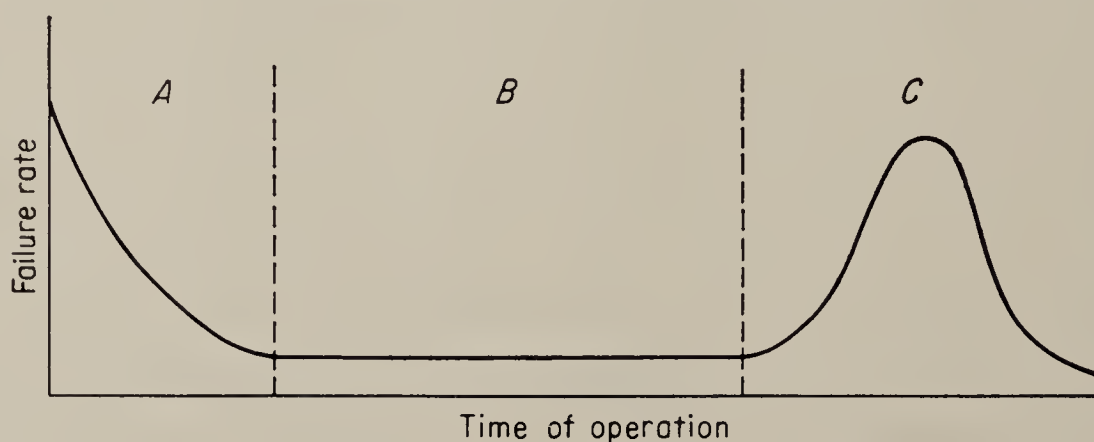


FIG. 1.46. Typical life history of a group of devices or components. Region *A* is the break-in region, *B* represents the operating region, and *C* is the wear-out region.

this region are few and random in nature and time of occurrence. Finally, region *C*, or the wear-out region, is a bell-shaped curve in a well-designed piece of equipment.

It has become a vogue in the guided-missile field to attempt to design equipment for short operating life. The argument goes that the unit need operate for only an hour or two at the most; thus, it is wasteful to

design for more. We can see the fallacy of this argument from Fig. 1.46. The break-in period blends into the wear-out region and there is essentially no period of trouble-free operation. We can categorically state that it is impossible to build a reliable, short-lived piece of equipment, and our missile program will be successful only when this is realized.

In general, the designer only shortens region B when he pushes components up to and beyond their design limits. It has been found, for example, that operating vacuum tubes at 120 per cent of their rated plate dissipation will result in about 200 per cent the failure rate of tubes operated at their design value. Several manufacturers have recently set up as standard-design procedure that vacuum tubes are to be operated at a maximum of 50 per cent of their rated plate dissipation. This has resulted in a 4:1 improvement in reliability.

The third cause of unreliability is operation under difficult environmental conditions. Extrapolation of present trends indicates that environmental specifications, such as ambient temperature, shock, and operation at low atmospheric pressures, will continue to become more severe. Derating of conventional components will soon fail to provide an operating margin. Extensive research on new materials and methods for component manufacture is imperative.

Finally, unreliability is due to improper maintenance. This may be the result of improperly trained technicians, lack of thought on the proper layout for ease of maintenance, or simply, equipment complexity. Usually all three causes are at work. One solution gaining favor is the use of modular packaging. Modular packaging of subunits of an assembly reduces repair to simply finding the troublesome module and replacing it. This is not only quicker and easier but is apparently more economical in the long run.

PROBLEMS

1.1. By means of resistance padding, construct the best approximation to the square law $y = 2x^2$, where x is shaft position in radians and y is resistance in ohms. Let $y_{\max} = 10,000$ and use a maximum of three segments.

1.2. It is desired to construct by voltage padding the response function shown in Fig. 1.47. Use four intermediate voltage taps.

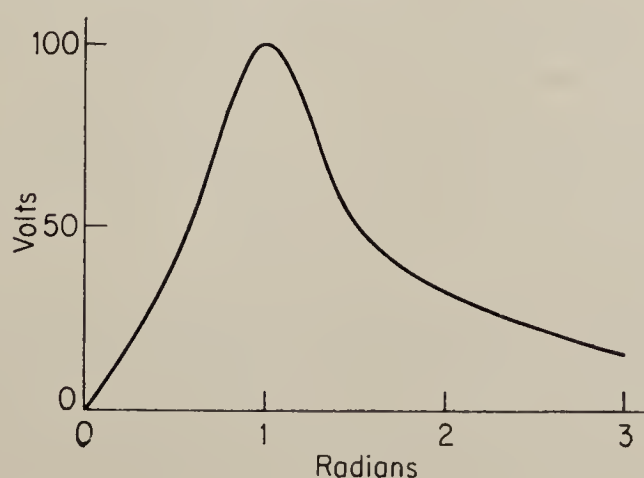


FIG. 1.47

1.3. By resistance, voltage padding approximate the cubic equation

$$y = x - x^2 + 0.5x^3$$

where x is shaft position in radians and y is the output voltage.

1.4. Draw the asymptotic α diagram for the functions:

$$\begin{aligned} a. & \frac{k(0.1s + 1)(0.05s + 1)}{s(0.01s + 1)(0.00001s^2 + 0.001s + 1)} \\ b. & \frac{s(5s + 1)(0.2s + 1)}{(10s + 1)(s + 1)(0.1s + 1)(0.02s + 1)} \end{aligned}$$

1.5. A network consists of two sections, H_1 and H_2 . The output impedance of H_1 is given by

$$Z_{o1} = \frac{k_1(0.5s + 1)(0.02s + 1)}{(2s + 1)(0.1s + 1)(0.01s + 1)}$$

and the input impedance of H_2 is

$$Z_{i2} = \frac{k_2(s + 1)(0.05s + 1)}{s(0.25s + 1)}$$

Find the ratio of k_1 to k_2 such that section H_2 does not load section H_1 .

1.6. Two lead networks (see Table 1.2) are to be cascaded to produce an over-all transfer function with a $+2$ slope in the asymptotic α diagram. The attenuation at low frequencies is the same for the two networks. Show that the ratio of impedance levels required to prevent loading between the two networks is a maximum when the two lead networks have the same break frequencies ω_1 and ω_2 .

1.7. Write down by inspection the impedance functions of the networks shown in Fig. 1.48.

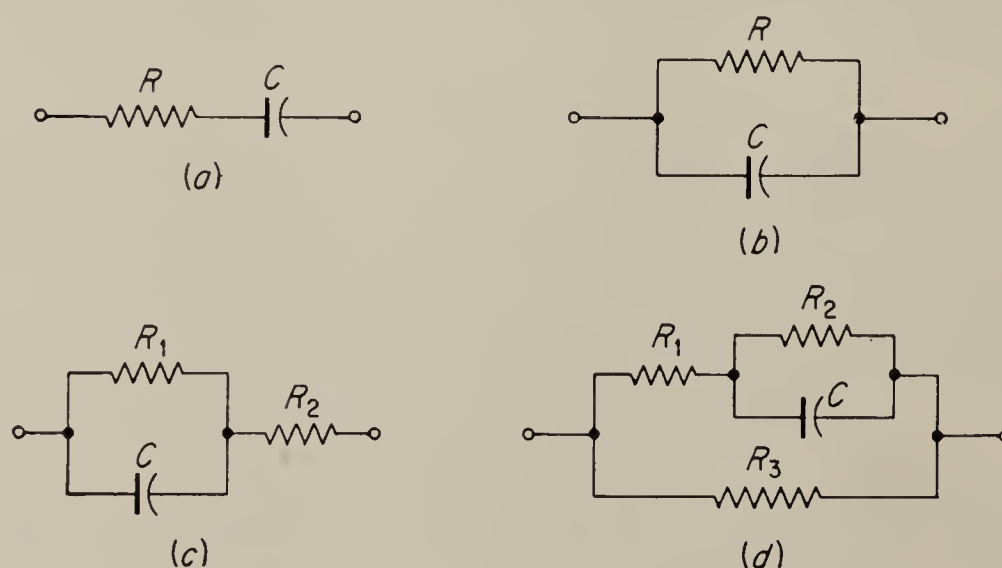


FIG. 1.48

1.8. For the driving-point impedances containing more than one capacitor shown in Fig. 1.49, write the impedance function by inspection by assuming that for any particular frequency only one capacitor is active. Also assume that the active zone for C_2 comes at a higher frequency than that of C_1 , etc.

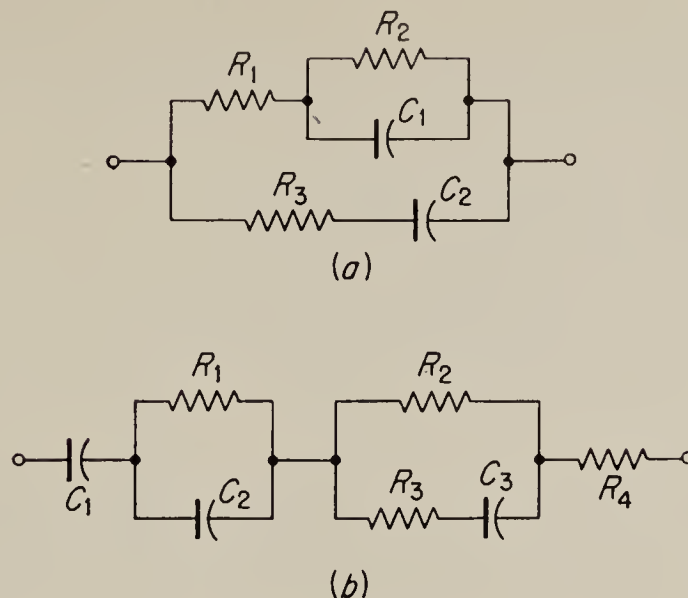


FIG. 1.49

1.9. Design a single L-section network by the approximate technique to synthesize the following transfer functions:

$$\begin{aligned}
 a. \quad H &= \frac{0.2s(0.05s + 1)}{(s + 1)(0.01s + 1)} \\
 b. \quad H &= \frac{(0.5s + 1)(0.05s + 1)^2}{(2s + 1)(0.2s + 1)(0.01s + 1)} \\
 c. \quad H &= h \frac{s(0.2s + 1)}{(s + 1)(0.05s + 1)(0.01s + 1)} \quad (h \text{ is to be as large as possible.}) \\
 d. \quad H &= \frac{(0.1s + 1)(0.025s + 1)}{(5s + 1)(0.01s + 1)(0.002s + 1)}
 \end{aligned}$$

1.10. Design the networks of Prob. 1.9a and b by the exact method described in Sec. 1.22.

1.11. Use two L sections to synthesize the following transfer functions. Use the approximate method described in Sec. 1.21 to obtain the L sections, and adjust the ratio of impedance levels to prevent loading.

$$\begin{aligned}
 a. \quad H &= \frac{1.25s^2(0.4s + 1)^2}{(10s + 1)(2s + 1)(0.1s + 1)^2} \\
 b. \quad H &= \frac{(0.05s + 1)^2(0.01s + 1)}{(s + 1)(0.005s + 1)^2} \\
 c. \quad H &= 0.1 \frac{(s + 1)(0.01s + 1)}{(0.1s + 1)^2} \\
 d. \quad H &= \frac{(0.0143s + 1)^3}{(s + 1)(0.1s + 1)(0.001s + 1)^2} \\
 e. \quad H &= h \frac{s^2(0.005s + 1)}{(0.1s + 1)^2(0.05s + 1)(0.02s + 1)} \quad (\text{Design for maximum } h.)
 \end{aligned}$$

1.12. Use the exact method of Sec. 1.22 to synthesize the individual L sections for the transfer functions given in Prob. 1.11. Compare the resulting designs to those obtained in Prob. 1.11.

1.13. Design a two-terminal impedance to have the following impedance function:

$$Z = 10,000 \frac{(0.02s + 1)(0.0025s + 1)}{(0.05s + 1)(0.01s + 1)}$$

Use both methods given in Sec. 1.24.

1.14. Use the exact synthesis method described in Secs. 1.23 to 1.25 to obtain the following transfer functions. Use the minimum number of L sections. In all cases design for maximum h .

$$a. H = h \frac{s^2(s + 500)(s + 2,000)^2}{(s + 10)^2(s + 50)(s + 10,000)^2}$$

$$b. H = h \frac{s(s + 10)(s + 20)(s + 40)}{(s + 1)(s + 3)(s + 5)(s + 80)}$$

$$c. H = h \frac{(s + 70)^3}{(s + 1)(s + 10)(s + 1,000)^2}$$

1.15. Find the input impedance and output impedance of the bridged-T network of Fig. 1.38. Assume the source impedance is zero and the load impedance infinite.

1.16. Repeat Prob. 1.15 for the twin-T network of Fig. 1.40, in which $R_1 = R_2 = \frac{1}{2}R_3$ and $C_1 = C_2 = 2C_3$.

1.17. A twin-T network for which $\alpha = \beta = n = 1$ is to be cascaded with a simple low-pass network (Table 1.2, second row). The pole in the low-pass network is to coincide with the null of the twin-T network. Find the ratio of impedance levels required to prevent appreciable loading between the two networks. Assume that the low-pass filter follows the twin T.

1.18. By the use of a bridged-T network together with one other L section, design a cascaded network to synthesize the following transfer function:

$$H = \frac{(2s + 1)(0.01s^2 + 0.05s + 1)(0.005s + 1)}{(10s + 1)(s + 1)(0.01s + 1)(0.001s + 1)}$$

Check that the two sections do not load each other appreciably.

1.19. Show that a twin-T network tuned to the carrier frequency of an a-c servo system has an effect on the carrier envelope that is similar to the effect of a perfect differentiating circuit on a low-frequency signal.

1.20. Design a bridged-T network to provide a 30° phase lead at a signal frequency of 6 cps for a servo using a 400-cps carrier. What is the effect of a 10 per cent shift in carrier frequency on the network characteristic?

1.21. An improperly tuned bridged-T network produces a phase shift of θ_u in the upper sideband and a phase shift θ_l in the lower sideband. Show that the effective phase lead of the envelope frequency is approximately $\frac{1}{2}(\theta_u - \theta_l)$. Also show that the phase shift of the carrier is approximately $\frac{1}{2}(\theta_u + \theta_l)$.

CHAPTER 2

D-C AMPLIFIERS

2.1. Introduction. D-c amplifiers are required as control amplifiers in a large class of electromechanical servomechanisms. In the following treatment it is assumed that the reader is familiar with the elementary rules of analysis and design of standard audio amplifiers, and the emphasis will, therefore, be primarily on points in which d-c amplifiers differ from the more conventional types.

A *d-c amplifier* is capable of amplifying signals at very low frequencies, including zero frequency or direct current. This means that the distinction between signal voltage and power-supply voltages, which in a conventional audio amplifier is obvious from the difference in frequency bands occupied, more or less completely disappears in the d-c amplifier and becomes largely a matter of definition. Thus, the output voltage is as much a function of the operating voltages as of the designated signal voltage and may have any value, even though the input voltage is zero. Since any power-supply-voltage variation or circuit-parameter variation will affect the output voltage, a d-c amplifier is subject to *drift*, i.e., a variation of the output independent of the designated input. Drift is undesirable, and its minimization poses one of the major problems in the design of d-c amplifiers.

Another difficulty arises from the fact that the low signal frequencies that must be handled make it impossible to use capacitors or transformers for interstage coupling or for any other separation of signal and power-supply or bias voltages. A d-c amplifier is characteristically *direct-coupled*. Hence the design of interstage coupling networks is another important problem facing the designer of d-c amplifiers.

Further requirements usually made of d-c amplifiers are similar to those made of audio amplifiers and require only brief mention. The gain of the amplifier should be stable and constant over as large a frequency range as possible. Any spurious output, usually referred to as *noise*, should be held to a minimum. It is usually desirable to have a high input impedance and a low output impedance (ideally infinity and zero, respectively). Ordinarily the output voltage should be zero when the input is zero, and an adjustment is usually provided somewhere in the circuit to accomplish this.

2.2. Drift in D-C Amplifiers. There are three major factors that cause drift in a d-c amplifier. These are:

1. Variations of power-supply voltages
2. Changes of components with temperature and age
3. Effects of heater-voltage variation

Only the third of these is assumed to be unfamiliar. Drift is caused by heater-voltage variations because changes in cathode temperature result in changes of the initial velocities of the electrons emitted, so that the electrode voltages required to maintain a given electron flow must change. It has been found that, if the plate current is small compared to the total emission from the cathode, this effect is essentially independent of the plate current. The heater-voltage effect is best expressed in terms of the amount by which the cathode voltage must change relative to the other electrode potentials in order to hold the plate current constant. For oxide-coated cathodes, this amounts to about 0.1 volt for a 10 per cent change in heater voltage about the normal value, and it is relatively independent of the tube type.¹

Various methods are used to counteract these effects. An obvious remedy is to regulate carefully all power supplies including heater supplies. High-quality components, wire-wound resistors, etc., will tend to minimize the effect of component variations. A number of cancellation methods have been devised to minimize the effect of cathode-temperature variation, and some of these will be described in detail below. These methods usually make use of the fact that the difference between the voltage changes required in two tubes to keep the plate current constant

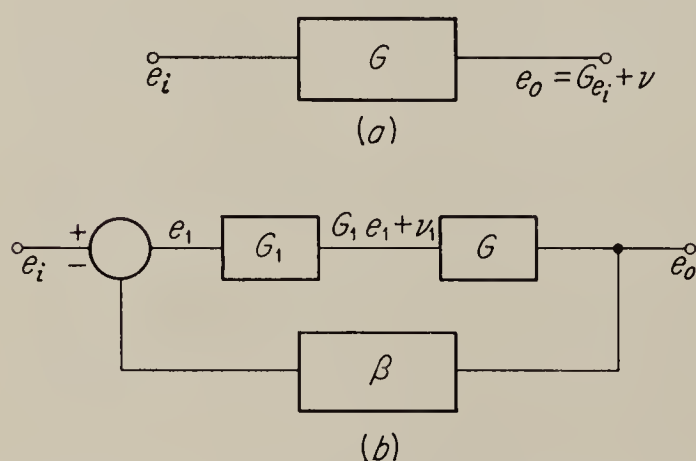


FIG. 2.1. Block diagrams of circuits used for the calculation of the effect of negative feedback on drift.

is much less than the change required in either tube alone. When they are applicable, push-pull circuits are very effective in reducing drift from all three causes, particularly if matched tubes and components are used. Negative feedback is usually ineffective, since it does not ordinarily affect the signal-to-noise ratio of an amplifier. This is demonstrated below.

2.3. Effect of Negative Feedback on Drift.

Suppose that the amplifier shown in Fig. 2.1a has a gain G and generates an internal noise having the value ν at the output. (Drift may be thought of as noise of very low

¹ An excellent discussion of this phenomenon can be found in Valley and Wallman, "Vacuum Tube Amplifiers," Radiation Laboratory Series, vol. 18, McGraw-Hill Book Company, Inc., New York, 1948, pp. 421-424.

frequency.) Then the output is

$$e_o = Ge_i + \nu \quad (2.1)$$

Suppose we attempt to improve this performance by using feedback. Since feedback alone will reduce the gain, we must add a preamplifier G_1 ahead of the original amplifier G , and, to be realistic, we must assume that the additional amplifier also contributes noise. Let this noise, measured at the output of G_1 , be ν_1 . The input voltage to G_1 is

$$e_1 = e_i - \beta e_o \quad (2.2)$$

Therefore $e_o = G_1 G e_1 + G \nu_1 + \nu$

$$\begin{aligned} &= G_1 G e_i - G_1 G \beta e_o + G \nu_1 + \nu \\ &= \frac{G_1 G}{1 + G_1 G \beta} e_i + \frac{G \nu_1}{1 + G_1 G \beta} + \frac{\nu}{1 + G_1 G \beta} \end{aligned} \quad (2.3)$$

For the feedback to have full effect, it is desirable that

$$G_1 G \beta \gg 1 \quad (2.4)$$

In this case, the denominator $1 + G_1 G \beta$ in Eq. (2.3) may be replaced approximately by $G_1 G \beta$, and Eq. (2.3) may be rewritten in the approximate form

$$e_o \approx \frac{1}{\beta} e_i + \frac{1}{G_1 \beta} \nu_1 + \frac{1}{G G_1 \beta} \nu \quad (2.5)$$

In order to have a valid comparison between the feedback amplifier and the original amplifier, we make the gain of both amplifiers the same; i.e., $1/\beta = G$. Then

$$e_o = G e_i + \frac{G}{G_1} \nu_1 + \frac{1}{G_1} \nu \quad (2.6)$$

We see that, although the original noise ν has been reduced by the gain of the preamplifier, G_1 , an additional noise term $(G/G_1)\nu_1$ has been added. If we assume that the amount of noise added by an amplifier is proportional to the amplifier gain, i.e., if $\nu_1 = (G_1/G)\nu$, then we see that the feedback amplifier has about the same noise output as the original amplifier.

The above argument does indicate, however, that if the preamplifier were noise-free, the noise output of the feedback amplifier would be considerably reduced. It also indicates that, in the design of a feedback amplifier, all care should be taken to make the first stages as noise-free as possible.

Getting back to the problem of drift, we shall find that it is possible to construct d-c amplifiers with zero drift and very high d-c gain. The gain does, however, drop sharply as the frequency is increased to values as high as 1 or 2 cps. If such an amplifier is used as a preamplifier in a feedback circuit like that of Fig. 2.1b, the drift introduced by the main ampli-

fier may be reduced to as small a value as is desired. The fact that the gain of the amplifier drops at higher frequencies is of no consequence, for as long as the inequality (2.4) is satisfied, the gain is given by $1/\beta$ and is essentially independent of the gain of the forward path of the feedback loop. This principle is employed in the chopper-stabilized d-c amplifier¹ described in detail later.

2.4. Analysis of Simple Triode Circuits. Before considering the operation and design of complete amplifiers, we must examine the behavior of single stages rather completely, with particular emphasis on the fact that they are to be used in d-c amplifiers. The quantitative effect on drift of the three factors listed in the last section, as well as questions of gain and terminal impedances, will be discussed.

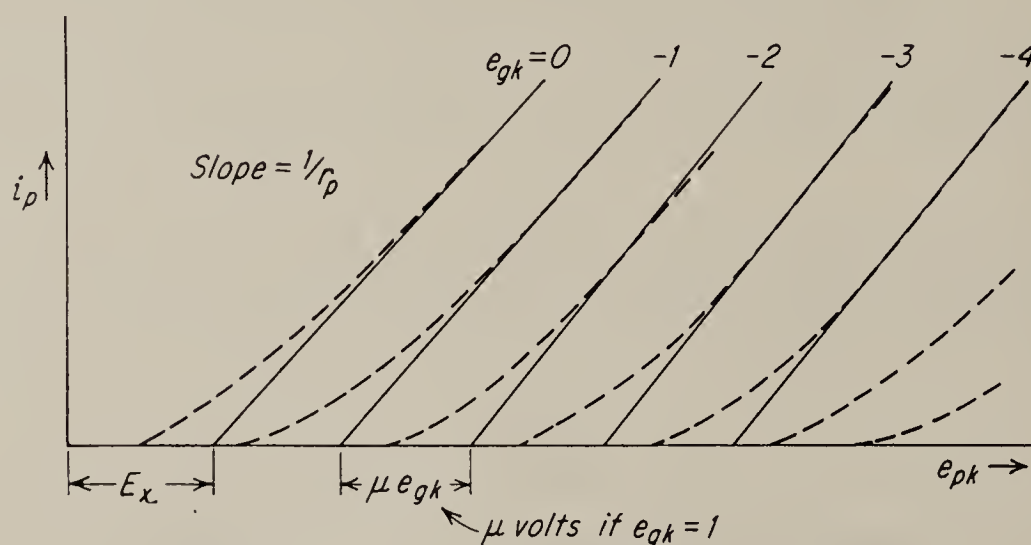


FIG. 2.2 Linear triode plate characteristics.

For purposes of analysis, the vacuum tube will be assumed completely linear with plate characteristics as shown in Fig. 2.2. The dotted lines represent the actual characteristics, while the solid lines represent the linear approximations.

The linear characteristics are parallel straight lines, equidistant for equal increments of grid voltage. The distance between the intercepts of two characteristics differing by a unit grid voltage is μ volts on the e_{pk} axis. The line for $e_{gk} = 0$ does not, in general, pass through the origin, and its intercept on the e_{pk} axis, usually positive, is E_x volts. It should be noted that these characteristics are derived from the published non-linear curves by drawing a tangent to the actual characteristic passing through the operating point. The distance between the lines (μe_{gk}) is determined by the average distance around the operating point. Thus the exact shape of the linearized characteristics depends very much on the operating point chosen.

By assuming these linear characteristics, the equation for the plate

¹ E. A. Goldberg, Stabilization of a Wide-band D-C Amplifier for Zero and Gain, *RCA Rev.*, June, 1950, pp. 296-300.

current can be found to be

$$i_p = \frac{1}{r_p} (e_{pk} + \mu e_{gk} - E_x) \quad (2.7)$$

where the voltages e_{pk} and e_{gk} are the plate and grid voltages, respectively, measured relative to the cathode. The tube may therefore be represented

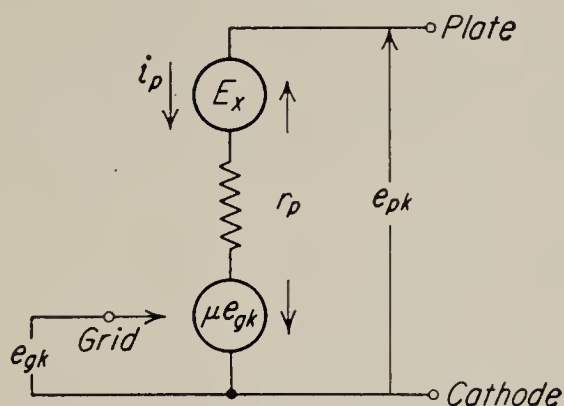


FIG. 2.3. Linear equivalent circuit.

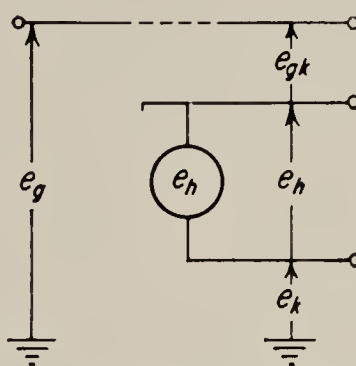


FIG. 2.4. Equivalent circuit of heater-voltage effect.

by the equivalent circuit shown in Fig. 2.3, where the arrows on the voltages represent the assumed positive sense.

This circuit is accurate only for normal cathode temperature, since it does not provide for the insertion of the cathode-temperature-variation effect. The initial velocity of the electrons results, approximately, in a “virtual” cathode voltage differing from the cathode voltage measured at the terminal. The effect of heater-voltage variation may, therefore, be most easily represented in the equivalent circuit by a voltage source e_h inserted between the cathode terminal and the cathode, as shown in Fig. 2.4. Hence the equivalent circuit of Fig. 2.3 may be generalized to include heater-voltage variation by considering the terminal marked “cathode” as the “virtual” cathode inside the tube and by connecting a voltage source e_h between this terminal and the outside connection. The voltage e_h may be defined as

$$e_h \triangleq - \left. \frac{\partial e_k}{\partial e_f} \right|_{i_p=c} \Delta e_f$$

where Δe_f is the change in filament voltage. This leads to the equivalent circuit of Fig. 2.5, which applies to all triodes. For this circuit the equation for the plate current becomes

$$i_p = \frac{1}{r_p} (e_p - E_x + \mu e_{gk} - e_h - e_k) \quad (2.8)$$

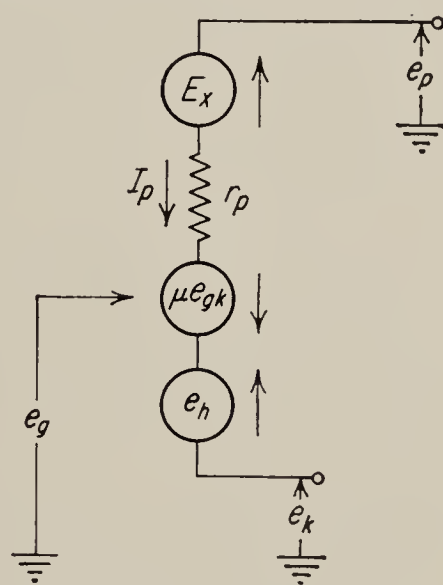


FIG. 2.5. Triode equivalent circuit with heater-voltage effect.

where e_{gk} is the voltage between the grid and the *virtual cathode*, as shown in Fig. 2.4.

2.5. The Cathode Follower. Cathode followers are used primarily as impedance changers, since they offer a very high input impedance and very low output impedance. The voltage gain is slightly less than unity. The actual circuit is shown in Fig. 2.6a and its equivalent in Fig. 2.6b.

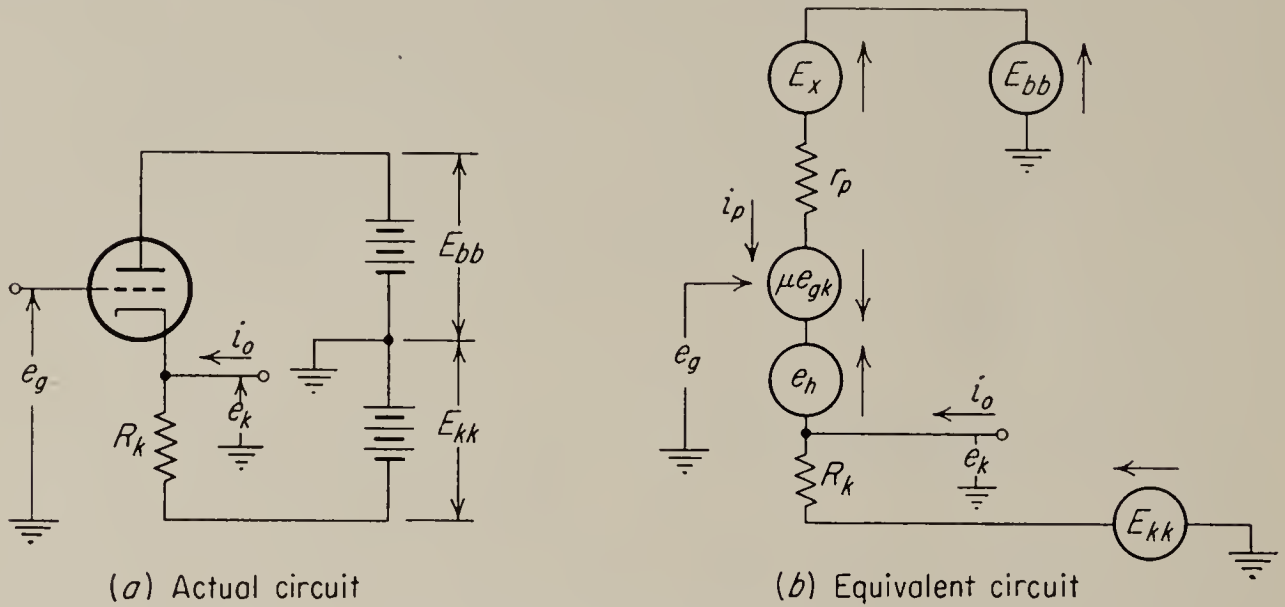


FIG. 2.6. Cathode follower.

The output voltage is e_k , measured between cathode and ground. The current i_o is the load current and is included for generality. It will be noted that the cathode is not returned to ground but to a negative voltage E_{kk} . This is necessary if the output voltage e_k of the cathode follower is to swing through negative as well as positive values of voltage.

The complete equation for this circuit may be derived by use of Kirchhoff's current law. The expression for the current leaving the node e_k is

$$(e_k - E_{kk}) \frac{1}{R_k} + (e_k + e_h - \mu e_{gk} + E_x - E_{bb}) \frac{1}{r_p} - i_o = 0 \quad (2.9)$$

But inspection of Fig. 2.4 reveals that

$$e_{gk} = (e_g - e_h - e_k) \quad (2.10)$$

$$\begin{aligned} \text{Hence } e_k \left(\frac{1}{R_k} + \frac{1}{r_p} + \frac{\mu}{r_p} \right) - E_{kk} \frac{1}{R_k} - (E_{bb} - E_x) \frac{1}{r_p} - e_g \frac{\mu}{r_p} \\ + e_h \left(\frac{1}{r_p} + \frac{\mu}{r_p} \right) - i_o = 0 \end{aligned} \quad (2.11)$$

$$\begin{aligned} \text{and } e_k &= \frac{e_g \frac{\mu}{r_p} + (E_{bb} - E_x) \frac{1}{r_p} + E_{kk} \frac{1}{R_k} - e_h \frac{1 + \mu}{r_p} + i_o}{\frac{1}{R_k} + \frac{1 + \mu}{r_p}} \\ &= \frac{\mu R_k e_g + R_k (E_{bb} - E_x) + r_p E_{kk} + R_k r_p i_o - (1 + \mu) R_k e_h}{R_k (\mu + 1) + r_p} \end{aligned} \quad (2.12)$$

In deriving this equation, no assumption other than linearity has been made. Hence this equation is correct for a-c or d-c, over the range of variation of the variables for which the assumption of linearity may be made to hold. In particular, the quiescent value of e_k may be found by setting $e_g = e_h = i_o = 0$ and inserting the proper values for E_{bb} , E_x , and E_{kk} .

Generally, however, we are interested in specializing this equation to obtain expressions for gain, output impedance, effect of power-supply-voltage and component variations on drift, etc. This is done by partial differentiation. Thus for the gain we have

$$G = \frac{\partial e_k}{\partial e_g} = \frac{\mu R_k}{(\mu + 1)R_k + r_p} \approx \frac{\mu}{\mu + 1} \approx 1 \quad (2.13)$$

The approximations made to obtain the final result are that $(\mu + 1)R_k \gg r_p$ and $\mu \gg 1$. For a typical triode (6SL7) common values are $\mu = 70$, $R_k = 200,000$ ohms, $r_p = 50,000$ ohms; hence these approximations are very good. Similarly we may find the output impedance

$$Z_o = \frac{\partial e_k}{\partial i_o} = \frac{R_k r_p}{(\mu + 1)R_k + r_p} \approx \frac{r_p}{\mu + 1} \approx \frac{1}{g_m} \quad (2.14)$$

Common values for the output impedance of a cathode follower are from 500 to 1,000 ohms, corresponding to a g_m of from 1,000 to 2,000 μ mhos. Note that the value of R_k does not appear in the approximate expression for the output impedance.

The drift due to changes in E_{bb} or E_{kk} is found from Eq. (2.12) to be

$$\frac{\partial e_k}{\partial E_{bb}} = \frac{R_k}{(\mu + 1)R_k + r_p} \approx \frac{1}{\mu + 1} \approx \frac{1}{\mu} \quad (2.15)$$

$$\text{and} \quad \frac{\partial e_k}{\partial E_{kk}} = \frac{r_p}{(\mu + 1)R_k + r_p} \approx \frac{r_p}{(\mu + 1)R_k} \approx \frac{1}{g_m R_k} \quad (2.16)$$

Note that drift from these two sources is rather small, especially if R_k is large. The drift due to heater variations is, however, large; it is given by

$$\frac{\partial e_k}{\partial e_h} = \frac{-(1 + \mu)R_k}{(1 + \mu)R_k + r_p} \approx -1 \quad (2.17)$$

Thus the drift due to heater-voltage variations is essentially equal to the change in e_h , or about 0.1 volt for every 10 per cent of heater-voltage change.

2.6. Input Impedance of the Cathode Follower. Since the cathode follower is used as an impedance-changing device, there is considerable interest in the value of the input impedance. Up to now, this has simply been assumed infinite. In practice, however, there is some leakage from the grid to the plate and cathode in any tube, which results in a finite

input impedance. Since the two leakage paths from the grid are in parallel, it is more convenient to consider the input admittance. We define leakage admittances between grid and plate and between grid and cathode as shown in Fig. 2.7. These admittances are assumed to be measured on a dead tube, i.e., on a tube for which $\mu = 0$.

The current i_g must be equal to

$$i_g = (e_g - E_{bb})Y_{gp} + (e_g - e_k)Y_{gk} \quad (2.18)$$

Therefore the input admittance is

$$\begin{aligned} Y_g &= \frac{\partial i_g}{\partial e_g} = Y_{gp} + \left(1 - \frac{\partial e_k}{\partial e_g}\right) Y_{gk} \\ &= Y_{gp} + \left(1 - \frac{\mu R_k}{(\mu + 1)R_k + r_p}\right) Y_{gk} \\ &= Y_{gp} + \frac{(R_k + r_p)Y_{gk}}{(\mu + 1)R_k + r_p} \approx Y_{gp} + \frac{Y_{gk}}{\mu + 1} \end{aligned} \quad (2.19)$$

where the assumption has been made that $R_k \gg r_p$, so that $R_k + r_p \approx R_k$.

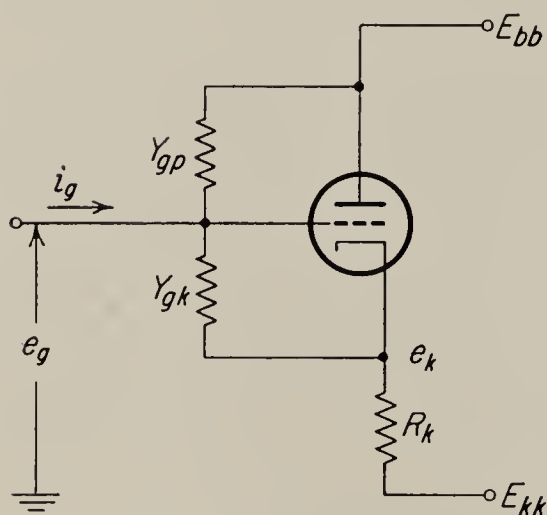


FIG. 2.7. Input-admittance paths in cathode follower.

Thus we see that, although the cathode follower has no effect on the grid-to-plate admittance (which is the same as the grid-to-ground admittance, since the plate is at “signal ground”), the grid-to-cathode admittance, measured when $\mu = 0$, is reduced approximately by the factor $1/(1 + \mu)$ when the tube is operating. This result should be compared with that obtained with a conventional grounded-cathode amplifier, where the input admittance is equal to the sum of the grid-to-cathode admittance plus the grid-to-plate admittance multiplied by

the gain of the stage.¹ This comparison shows the cathode follower to have a much lower input admittance than an amplifying stage.

2.7. Drift due to Component Variation. Equation (2.12) can also be used for computing the drift caused by component variations. Thus, for instance,

$$\begin{aligned} \frac{\partial e_k}{\partial R_k} &= \frac{r_p[\mu e_g + (E_{bb} - E_x) + r_p i_o - (\mu + 1)(E_{kk} + e_h)]}{[R_k(\mu + 1) + r_p]^2} \\ &\approx \frac{r_p[\mu e_g + (E_{bb} - E_x) + r_p i_o - (\mu + 1)E_{kk}]}{[(\mu + 1)R_k + r_p]^2} \end{aligned} \quad (2.20)$$

¹ Reich, “Theory and Applications of Electron Tubes,” McGraw-Hill Book Company, Inc., New York, 1944, pp. 93–96.

since $e_h \ll E_{kk}$. Also we have

$$\frac{\partial e_k}{\partial \mu} = \frac{R_k[(R_k + r_p)e_g - R_k(E_{bb} - E_x) - R_k r_p i_o - r_p(E_{kk} + e_h)]}{[(\mu + 1)R_k + r_p]^2} \quad (2.21)$$

These relations and similar ones for other component variations are useful, not only for the computation of drift, but also to determine the maximum range of deviation of the output voltage as a result of component tolerances. If such a range is known, it becomes a relatively simple matter to specify the range of balancing potentiometers required in the amplifier. Note that, in the application of formulas for drift due to component variations, all voltages and currents involved in the circuit must in general be known.

2.8. Cathode Follower with Plate-load Resistor. Occasionally, a plate-load resistor R_L is inserted between the plate and the positive supply voltage of a cathode follower. This occurs particularly in situations where the tube operates simultaneously as an amplifier stage and a cathode follower. The analysis derived thus far applies to cases of this sort almost without change; all that is required is the replacement of r_p in all the relations by $r_p + R_L$. This does, however, have the result, particularly if R_L is large, that a number of the approximations made to simplify the final expressions are no longer completely valid.

2.9. Equivalent Circuit for the Cathode Follower. In the analysis of complicated circuits involving cathode followers, it is sometimes convenient to make use of an equivalent circuit. Such a circuit may, of

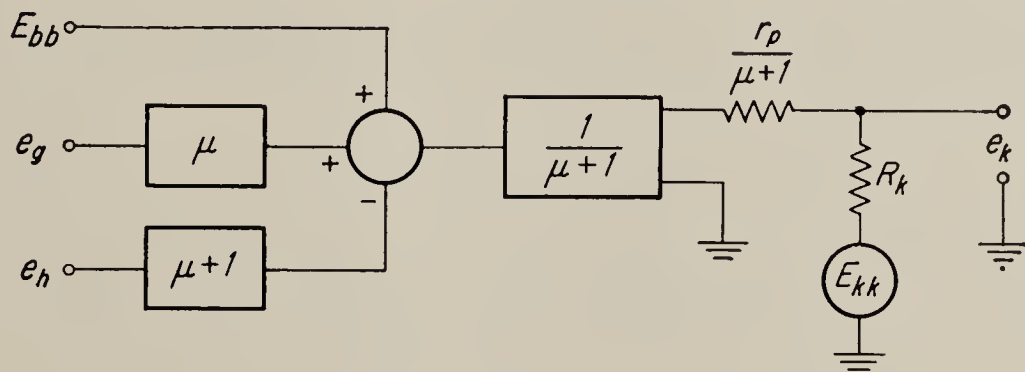


FIG. 2.8. Cathode-follower equivalent circuit.

course, take any number of forms. It is merely required that the expressions for gain, drift, and impedance derived from the equivalent circuit correspond to those derived previously. The circuit shown in Fig. 2.8 is suggested for situations where only the gain, output impedance, and drift rate due to E_{bb} , E_{kk} , and e_h are of interest. This circuit does *not* give the correct answer in problems involving component variations. The boxes marked μ , $\mu + 1$, and $1/(\mu + 1)$ are perfect amplifiers with zero output impedance. The input and output voltages, as well as the output current, are to be thought of as voltage and current variations about some fixed, unspecified quiescent value. For this reason it is not necessary to

include the intercept voltage E_x , which is always assumed to be constant. Note that the equivalent circuit remains valid even if R_k is removed, i.e., becomes infinite. This fact has occasional practical significance.

2.10. The Triode Amplifier with Cathode Bias. Before deriving expressions for the performance of the triode amplifier, it is convenient to

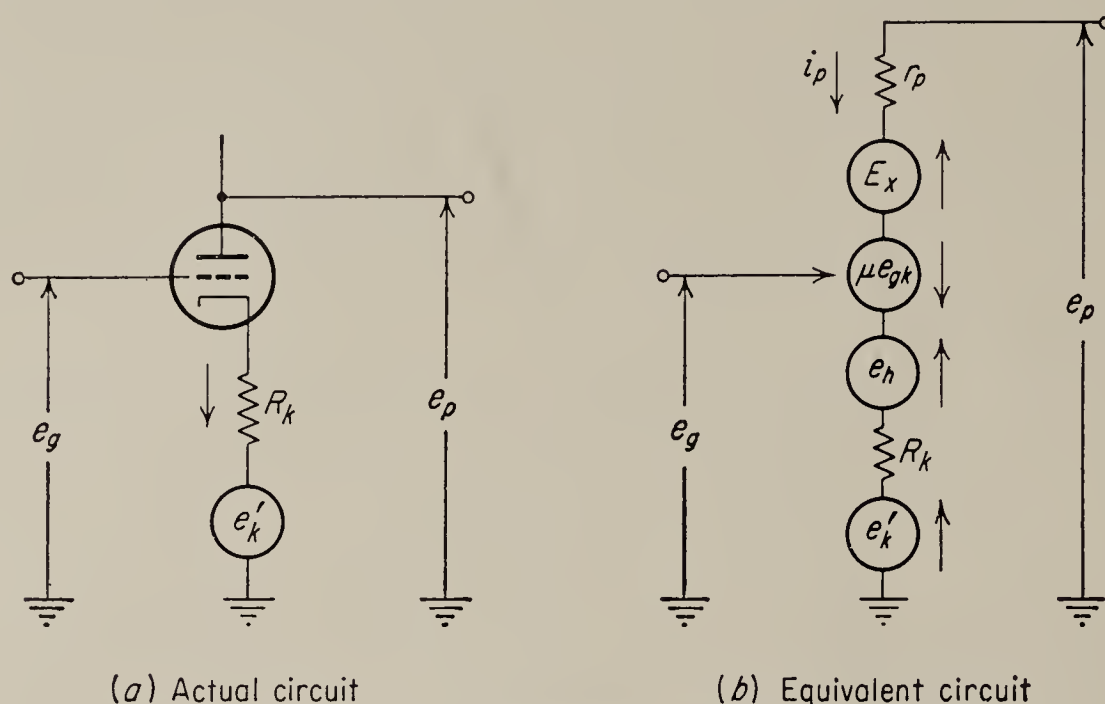


FIG. 2.9. Triode amplifier with cathode bias.

determine the effect of the cathode resistor. Thus, for the circuit shown in Fig. 2.9, the plate current is

$$i_p = \frac{e_p - E_x + \mu e_{gk} - e_h - e'_k}{r_p + R_k} \quad (2.22)$$

(The voltage e'_k is a cathode-signal voltage inserted for generality.) Since

$$e_{gk} = e_g - e_h - R_k i_p - e'_k \quad (2.23)$$

the current becomes

$$\begin{aligned} i_p &= \frac{e_p - E_x + \mu e_g - (\mu + 1)(e_h + e'_k) - \mu R_k i_p}{r_p + R_k} \\ &= \frac{e_p - (E_x + e_h + e'_k) + \mu(e_g - e_h - e'_k)}{r'_p} \end{aligned} \quad (2.24)$$

where

$$r'_p = r_p + (\mu + 1)R_k \quad (2.25)$$

If $R_k = 0$ in Eq. (2.25), it is obvious that, for a tube without a cathode resistor, the plate current is given by the same expression, except that $r'_p = r_p$. Hence we arrive at a result of general importance: Whenever a tube is viewed only from the three terminals—plate, grid, and lower end of cathode resistor—it may be replaced by an equivalent tube whose plate resistance is $r'_p = r_p + (\mu + 1)R_k$ and whose other characteristics are unchanged.

We make use of this fact to derive the performance equations for a

triode amplifier. The equivalent circuit is drawn in Fig. 2.10 in accordance with the results obtained in Eq. (2.24), and a node equation is written at node e_p , the output node. The voltage e'_k is retained for generality.

We have

$$\frac{e_p - E_{bb}}{R_L} + \frac{e_p - E_x + \mu e_g - (e_h + e'_k)(1 + \mu)}{r'_p} - i_o = 0 \quad (2.26)$$

$$\text{or } e_p = \frac{E_{bb}r'_p + E_xR_L - \mu R_L e_g + (\mu + 1)R_L(e_h + e'_k) + R_L r'_p i_o}{R_L + r'_p} \quad (2.27)$$

To find the gain, we differentiate e_p with respect to e_g and obtain the familiar result

$$G = \frac{\partial e_p}{\partial e_g} = \frac{-\mu R_L}{R_L + r'_p} = \frac{-\mu R_L}{R_L + r_p + (\mu + 1)R_k} \quad (2.28)$$

The output impedance is

$$Z_o = \frac{\partial e_p}{\partial i_o} = \frac{R_L r'_p}{R_L + r'_p} = \frac{R_L[r_p + (\mu + 1)R_k]}{R_L + r_p + (\mu + 1)R_k} \quad (2.29)$$

This could have been found also by inspection of Fig. 2.10, since all voltage sources in that figure are assumed to have zero impedance.

The drift due to E_{bb} is

$$\frac{\partial e_p}{\partial E_{bb}} = \frac{r'_p}{R_L + r'_p} = \frac{r_p + (\mu + 1)R_k}{R_L + r_p + (\mu + 1)R_k} \quad (2.30)$$

This again could have been found by inspection. The drift due to heater-voltage variation is obtained by differentiating e_p with respect to e_h :

$$\frac{\partial e_p}{\partial e_h} = \frac{(\mu + 1)R_L}{R_L + r_p + (\mu + 1)R_k} \quad (2.31)$$

Here we find again, as with the cathode follower, that filament-voltage variation causes a very large amount of drift, since the effective cathode voltage e_h acts on the tube in approximately the same way as the grid voltage, which is the desired signal.

Drifts due to variations of the components may be obtained by differentiating Eq. (2.27) with respect to the desired component. Specific computations are left to the reader.

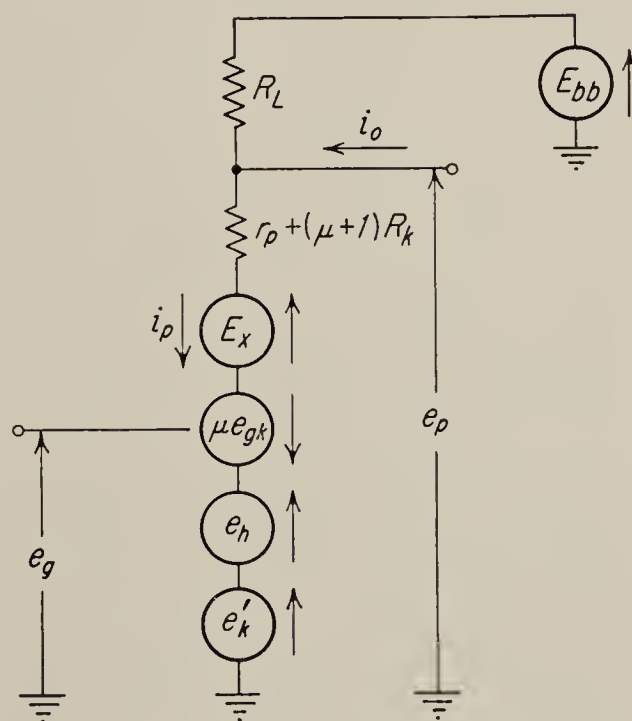


FIG. 2.10. Equivalent circuit of triode amplifier with cathode resistor.

If the tube is operated as a grounded-grid amplifier with cathode input, the voltage e'_k will become the signal voltage. The gain of the tube for this type of operation may be computed from Eq. (2.27), and the result is identical with Eq. (2.31):

$$\frac{\partial e_p}{\partial e'_k} = \frac{(\mu + 1)R_L}{R_L + r_p + (\mu + 1)R_k} \quad (2.32)$$

Comparison of this equation with (2.28) shows that the gain for the two methods of signal introduction is about the same, the gain for cathode input being $(\mu + 1)/\mu$ times that for grid input. More important is the fact that the sign change associated with amplification when the signal is applied to the grid does not occur when the signal is applied to the cathode. This fact is of importance when it is desired to apply negative feedback around an amplifier.

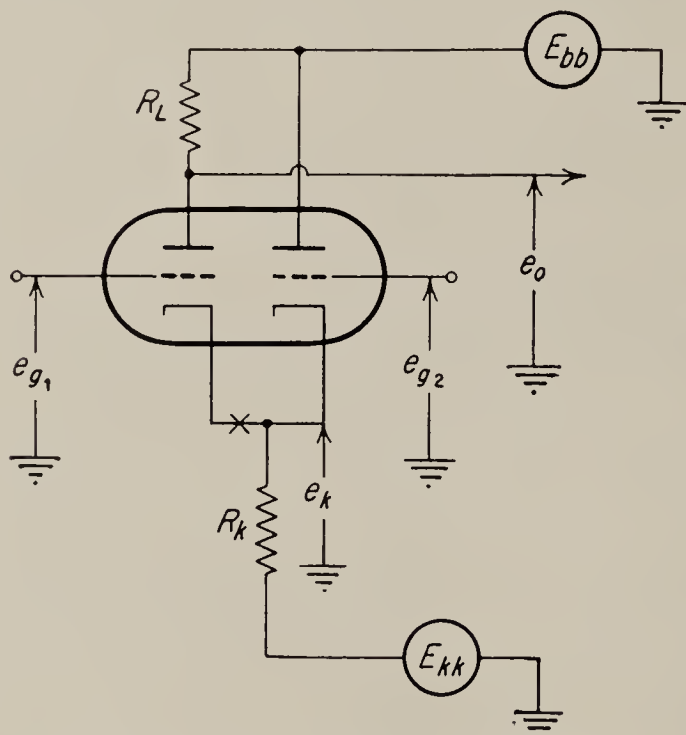


FIG. 2.11. The difference amplifier.

One disadvantage of applying the signal to the cathode is that the input impedance is very low. The input impedance may be deduced from one of the results obtained for the cathode follower; i.e., it is the sum of R_k and the output impedance of a cathode follower with a plate-load resistor R_L and an infinite cathode-load resistor. Figure 2.9a shows the circuit; the impedance in question is that "seen" by e'_k . Hence

$$Z_i = R_k + \frac{r_p + R_L}{\mu + 1} \quad (2.33)$$

This is usually much less than the input impedance at the grid.

2.11. The Difference Amplifier. We have seen that one of the most serious causes of drift is the variation of heater voltage, with its influence on the effective cathode voltage. A number of methods have been devised to counteract this effect, one of the most useful being the difference amplifier, shown in Fig. 2.11. The two triodes may be in separate envelopes; however, the two sections in a single envelope are usually better matched and will, therefore, give better results than two triodes picked at random.

To analyze the circuit, we suppose that the cathode of the left-hand section is disconnected at the point marked x in the diagram. The right half of the tube is then a standard cathode follower, and if for the moment we consider only the effect of the grid voltage on the output, we have

$$e_k = \frac{e_{g2}\mu_2 R_k}{(\mu_2 + 1)R_k + r_{p2}} \quad (2.34)$$

The subscript 2's refer to the right-hand triode. Also the impedance to ground seen by the left-half triode is the output impedance of the cathode follower. It is given by

$$Z = \frac{r_{p2}R_k}{(1 + \mu_2)R_k + r_{p2}} \quad (2.35)$$

Hence the circuit may be redrawn in the equivalent form of Fig. 2.12. The output voltage e_o for this circuit as a function of e_{g1} and e_{g2} is given by the combination of Eqs. (2.28) and (2.32), with R_k in both equations replaced by Z of Eq. (2.35). Thus,

$$e_o = \frac{-\mu_1 R_L e_{g1} + (\mu_1 + 1)R_L \frac{\mu_2 R_k e_{g2}}{(\mu_2 + 1)R_k + r_{p2}}}{R_L + r_{p1} + (\mu_1 + 1) \frac{r_{p2}R_k}{(\mu_2 + 1)R_k + r_{p2}}} \quad (2.36)$$

The subscript 1 refers to the left-hand triode. This equation may be simplified by the assumptions that $(\mu_2 + 1)R_k \gg r_{p2}$ and that $\mu_1 = \mu_2$ and $r_{p1} = r_{p2}$; hence

$$e_o \approx \frac{\mu R_L (e_{g2} - e_{g1})}{R_L + 2r_p} \quad (2.37)$$

It should be noted that, since the plate circuits of the two sections are different, one would expect the triodes to have different operating points and, therefore, different values for μ and r_p . Thus, although Eq. (2.37) is usually remembered because of its simplicity, it is not accurate. It is, however, correct to conclude that the output voltage is a function of the difference of the two grid voltages. In many circuits this property is of considerable use. For instance, one of the grids may be connected to a variable bias supply for zero adjustment of an amplifier, or a feedback voltage may be introduced into an amplifier through one of the grids, etc.

The effect of heater-voltage variations may be computed by inserting voltages e_{h1} and e_{h2} between the "actual" cathodes and the cathode terminals of the left- and right-hand sections, respectively. An analysis similar to that carried out for e_{g1} and e_{g2} results in

$$e_o = \frac{(\mu_1 + 1)R_L e_{h1} - (\mu_1 + 1)R_L \frac{(\mu_2 + 1)R_k e_{h2}}{(\mu_2 + 1)R_k + r_{p2}}}{R_L + r_{p1} + (\mu_1 + 1) \frac{r_{p2}R_k}{r_{p2} + R_k(\mu_2 + 1)}} \approx \frac{(\mu + 1)R_L (e_{h1} - e_{h2})}{R_L + 2r_p} \quad (2.38)$$

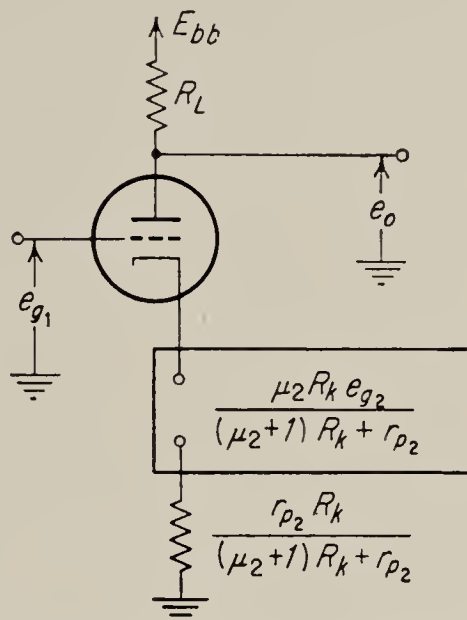


FIG. 2.12. Equivalent form for difference amplifier.

where the simplification is again due to the assumption of large R_k and equal μ and r_p in the two tube halves. The importance of this result lies in the fact that it is the *difference* between the initial-velocity voltages e_{h1} and e_{h2} that is amplified by the stage. This difference is, on the average, at least ten times smaller than either one of the two voltages, particularly if two triode sections in the same envelope are used. Thus this circuit, while giving substantially the same amplification as a single triode stage, results in a reduction of drift by a factor of ten or more.

The drift performance of this circuit for variations of the positive and negative supply is not quite so favorable but is still quite acceptable. The computation for E_{bb} can be carried out along the same lines as indicated for the grid voltages and offers no difficulty. The computation for drift due to E_{kk} variations is best carried out by substituting the equivalent circuit of Fig. 2.8 for the cathode-follower section of the difference amplifier. Both of these computations are left to the reader.

2.12. The Miller Circuit. Although the difference amplifier makes possible a large reduction in the effect of heater-voltage variations, it will

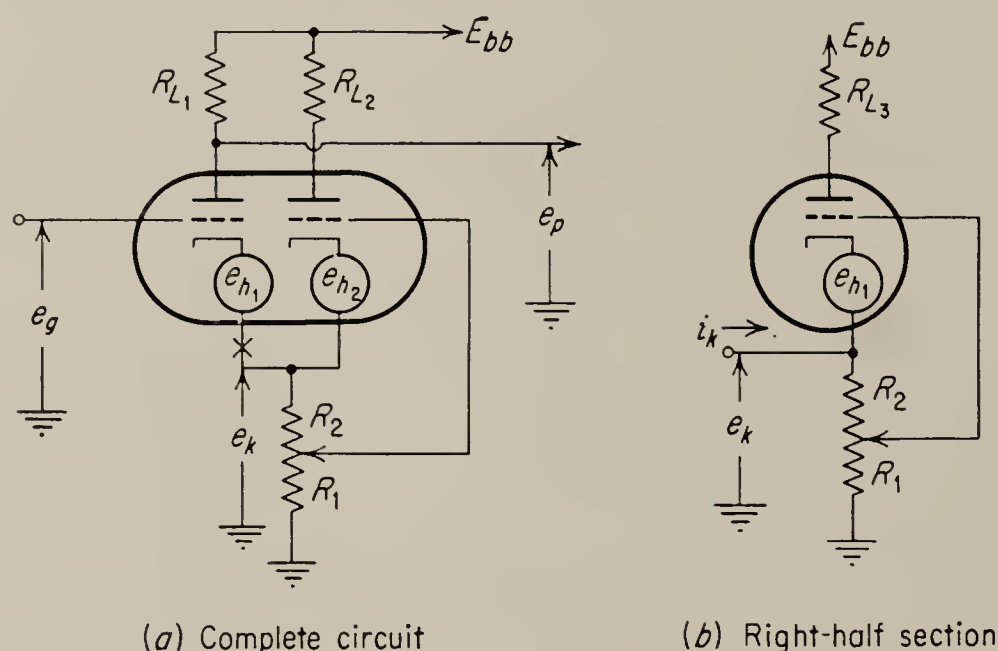


FIG. 2.13. The Miller circuit.

not, in general, reduce the drift from this cause completely to zero. Since no two tubes have exactly the same characteristics, a circuit which is to be insensitive to heater-voltage variation must contain an adjustment to accommodate different tubes. Such a circuit was described by Miller¹ and takes the form shown in Fig. 2.13.

The analysis proceeds as for the difference amplifier. We assume the existence of the heater-effect voltages e_{h1} and e_{h2} in the two sections of the tube and, for the moment, assume the left-hand section disconnected at point x . The right-hand section may then be considered as a cathode

¹ S. E. Miller, Sensitive D-C Amplifier with A-C Operation, *Electronics*, November, 1941, p. 27.

follower with a plate-load resistance. Its input voltage is

$$e_{g2} = \frac{R_1}{R_1 + R_2} e_k = \frac{R_1}{R_k} e_k \quad (2.39)$$

(see Fig. 2.13b). The voltage at the break x may then be deduced by the superposition of Eqs. (2.13) and (2.17). It is necessary, however, to replace r_p by $r_{p2} + R_{L2}$ in accordance with the discussion in Sec. 2.8. Also e_g in Eq. (2.13) must be replaced by Eq. (2.39). The result is

$$\begin{aligned} e_k &= \frac{-(\mu_2 + 1)R_k e_{h2} + \mu_2 R_k (R_1/R_k) e_k}{R_k(\mu_2 + 1) + r_{p2} + R_{L2}} \\ &= \frac{-(\mu_2 + 1)R_k e_{h2}}{R_k(\mu_2 + 1) + r_{p2} + R_{L2} - \mu_2 R_1} \end{aligned} \quad (2.40)$$

where $R_k = R_1 + R_2$.

We may now redraw Fig. 2.13a as Fig. 2.14, where R_o is the output impedance of the cathode follower of Fig. 2.13b. It may be seen from this figure that, as long as

$$e_{h1} = e_{h2} \frac{(\mu_2 + 1)R_k}{R_k(\mu_2 + 1) + r_{p2} + R_{L2} - \mu_2 R_1} \quad (2.41)$$

the net voltage in the cathode circuit of the left section is zero, and the heater-voltage effect is completely canceled. Since the factor multiplying e_{h2} is an adjustable constant, because of the presence of R_1 , the circuit can be made insensitive to heater-variation effects as long as there is a constant ratio between e_{h1} and e_{h2} .

This constant ratio exists for a fairly large range of filament voltages around the proper operating value. Accordingly, if we let

$$e_{h1} = a e_{h2} \quad (2.42)$$

we can solve for the value of R_1 that results in zero drift. The result is

$$R_1 = R_k \left(1 - \frac{1}{a}\right) \left(1 + \frac{1}{\mu_2}\right) + \frac{r_{p2} + R_{L2}}{\mu_2} \quad (2.43)$$

$$\approx R_k \left(1 - \frac{1}{a}\right) + \frac{r_{p2} + R_{L2}}{\mu_2} \quad (2.44)$$

The chief function of R_{L2} is to prevent overloading of the right-hand section of the tube, and it must be large enough for this purpose. On the other hand, excessively large values of R_{L2} result in a decrease in gain for

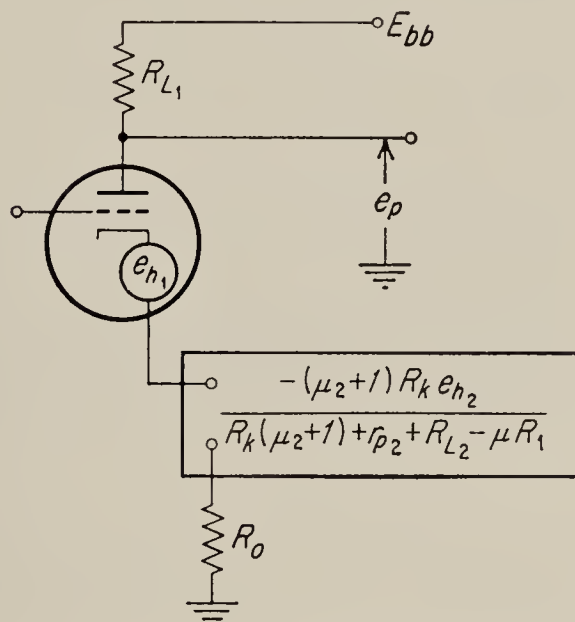


FIG. 2.14. Equivalent circuit for the Miller circuit.

the circuit [see Eq. (2.49) below]; hence it should not be larger than necessary.

In practice it is undesirable to make the cathode resistor, $R_1 + R_2 = R_k$, constant, as implied in the previous analysis, since any adjustment of the variable tap affects the bias voltage of both grids, making it very difficult to prevent either one or the other of the two tube sections from cutting off. A more practical form of the circuit is shown in Fig. 2.15. R_3 and R_4 together form the effective plate load of the right half section and serve to

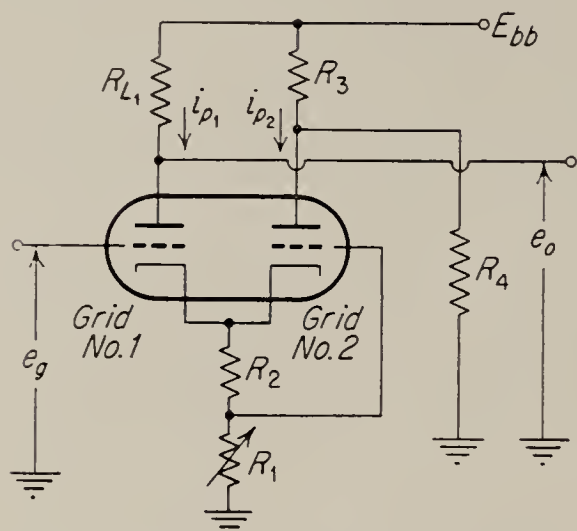


FIG. 2.15. Practical form of Miller circuit.

reduce the effective E_{bb} seen by this section. Thus i_{p2} is kept small and does not produce a large bias through the common cathode resistor. R_2 is constant and small, just enough to keep grid 2 slightly negative with respect to its cathode. (A small amount of grid current in the right-half section is actually of little consequence.) R_{L1} should be large, both for maximum gain and for wide operating range, since the adjustment of R_1 required for drift compensation does move the operating point of the left-half section over fairly wide

limits. (It has only a small effect on the right half.)

The condition for perfect balance now becomes

$$R_1 = (R_1 + R_2) \left(1 - \frac{1}{a}\right) \left(1 + \frac{1}{\mu_2}\right) + \frac{r_{p2} + R'_{L2}}{\mu_2} \quad (2.45)$$

$$\approx R_2(a - 1) + \frac{a}{\mu_2} (r_{p2} + R'_{L2})$$

where

$$R'_{L2} = \frac{R_3 R_4}{R_3 + R_4}$$

The reader may convince himself, by using typical values in the circuit of Fig. 2.15, that reasonable variations of a can be compensated for without causing tube cutoff.

In order to find the gain of the circuit or its sensitivity to changes in the positive supply voltage, it is necessary first to compute the effective cathode resistor seen by the amplifying section. This is best done by writing an expression for the voltage change e_k as a function of a current i_k flowing into the right-half section (see Fig. 2.13b). Using Eqs. (2.13), (2.14), and (2.39) and the fact that r_p must be replaced by $r_{p2} + R'_{L2}$, we have

$$e_k = \frac{i_k R_k (r_{p2} + R'_{L2})}{R_k (\mu_2 + 1) + r_{p2} + R'_{L2}} + \frac{\mu_2 R_k (R_1 / R_k) e_k}{R_k (\mu_2 + 1) + r_{p2} + R'_{L2}}$$

$$= \frac{i_k R_k (r_{p2} + R'_{L2})}{R_k (\mu_2 + 1) + r_{p2} + R'_{L2} - \mu_2 R_1} \quad (2.46)$$

Hence the effective cathode resistor of the left-half section (R_o of Fig. 2.14) is

$$R_o = \frac{e_k}{i_k} = \frac{R_k(r_{p2} + R'_{L2})}{R_k(\mu_2 + 1) + r_{p2} + R'_{L2} - \mu_2 R_1} \quad (2.47)$$

If the circuit is properly adjusted to eliminate drift due to heater-voltage variations, we may use Eq. (2.43) to eliminate R_1 with the simple result that

$$R_o = \frac{a(r_{p2} + R'_{L2})}{\mu_2 + 1} \quad (2.48)$$

where a is defined as in Eq. (2.42). Now the expression for gain found previously for the triode amplifier with cathode bias [Eq. (2.28)] may be used, with the result:

$$\begin{aligned} G = \frac{\partial e_p}{\partial e_g} &= - \frac{\mu_1 R_{L1}}{R_{L1} + r_{p1} + (\mu_1 + 1)(r_{p2} + R'_{L2})a/(\mu_2 + 1)} \\ &\approx - \frac{\mu_1 R_{L1}}{R_{L1} + r_{p1} + a(r_{p2} + R'_{L2})} \quad \text{if } \mu_1 = \mu_2 \end{aligned} \quad (2.49)$$

The computation of the drift due to changes in E_{bb} is slightly more involved, since the drift is due both to the direct effect of E_{bb} on the plate voltage [as in Eq. (2.30)] and to its “indirect” effect on the cathode voltage e_k of the right-half section and thereby on the left-hand plate voltage. The details of this computation are left to the reader. For the properly adjusted circuit the result is

$$\frac{\partial e_p}{\partial E_{bb}} \approx \frac{r_{p1} + a(R_{L1} + r_{p2} + R'_{L2})}{r_{p1} + R_{L1} + a(r_{p2} + R'_{L2})} \quad (2.50)$$

where we again assume that $\mu_1 = \mu_2$. It is clear that, for $a = 1$, Eq. (2.50) becomes $\Delta e_p = \Delta E_{bb}$. Hence the Miller circuit does not compensate for changes in the positive supply voltage very well, and for best results the positive supply must be closely regulated.

2.13. Phase Inverters. Whenever it is possible, push-pull amplifiers are used because of their relatively drift-free operation. However, if the input signal is with respect to the ground, “single-ended,” a phase inverter is required to convert the signal into the “double-ended” form required in push-pull circuits. A number of fairly satisfactory circuits are available for this purpose.

2.14. The Cathode-follower Inverter. The simplest form of inverter is the cathode-follower type shown in Fig. 2.16. For the output voltage e_{o1} the circuit looks like a triode amplifier with cathode

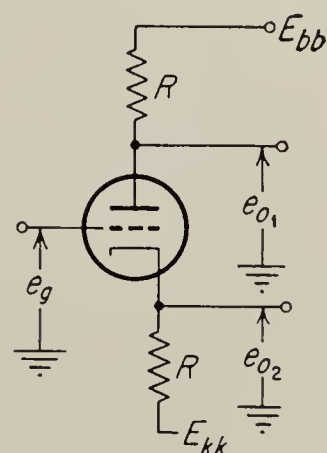


FIG. 2.16. The cathode-follower inverter.

bias; hence Eq. (2.28) applies and results in

$$\frac{e_{o1}}{e_g} = - \frac{\mu R}{(\mu + 2)R + r_p} \approx -1 \quad (2.51)$$

The output voltage e_{o2} is obtained by treating the circuit as a cathode follower with plate-load resistance R ; hence Eq. (2.13) may be used and gives

$$\frac{e_{o2}}{e_g} = \frac{\mu R}{(\mu + 2)R + r_p} \approx 1 \quad (2.52)$$

Hence we have here exactly that $e_{o1}/e_{o2} = -1$, the ideal inverter action. Unfortunately, if e_g is near ground potential, e_{o1} is at a relatively high positive d-c level, whereas e_{o2} is near ground. This is no disadvantage in an a-c amplifier, but since most d-c amplifier coupling networks introduce some signal loss, the voltage e_{o1} will generally be attenuated somewhat before it can be applied to a push-pull stage. This effect may, of course, be compensated by making the plate-load resistor larger than the cathode

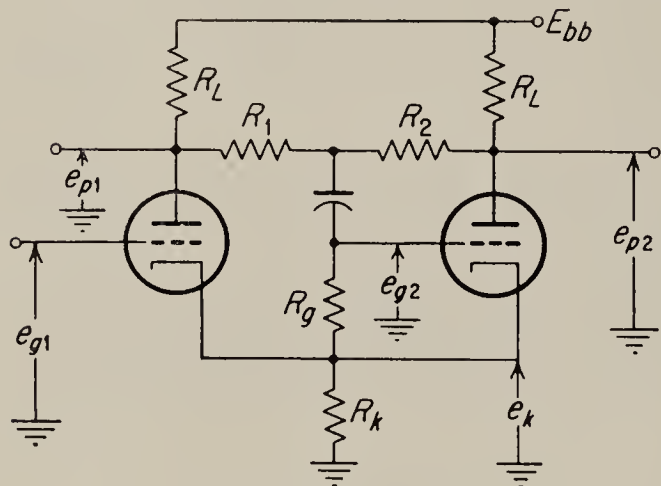


FIG. 2.17. The paraphase inverter.

any of the d-c coupling networks described below in Sec. 2.17 will make this circuit applicable to d-c operation.

The circuit analysis is simplified if it is assumed that

1. The two tubes are identical; i.e., r_p and μ are the same.
2. $R_1 = R_2 \gg R_L$.
3. $R_g \gg R_1$.
4. The frequency is high enough for the impedance of the coupling capacitor to be negligible.
5. The inverter action is sufficiently good that the current in the cathode resistor R_k is constant; hence the cathode voltage e_k is constant.

The following three equations may then be written for this circuit:

$$e_{p1} = \frac{-\mu R_L}{R_L + r_p} e_{g1} \quad (2.53)$$

$$e_{p2} = \frac{-\mu R_L}{R_L + r_p} e_{g2} \quad (2.54)$$

$$e_{g2} = \frac{1}{2}(e_{p1} + e_{p2}) \quad (2.55)$$

resistor, but then the circuit loses some of its simplicity. Another disadvantage of this circuit is that it has no gain. Its major advantage is its simplicity and the fact that only one tube is required.

2.15. The Paraphase, or Balanced, Inverter. This inverter takes a form similar to that shown in Fig. 2.17, the common circuit used in a-c amplifiers. Replacing the capacitor by

Substitution of (2.55) in (2.54) yields

$$e_{p2} = \frac{-\mu R_L}{R_L + r_p} \frac{e_{p1} + e_{p2}}{2} = \frac{-\mu R_L}{\mu R_L + 2R_L + 2r_p} e_{p1} \quad (2.56)$$

If μ is very large, this becomes approximately $e_{p2} = -e_{p1}$, a typical value of $-e_{p2}/e_{p1}$ being of the order of 0.95. Thus the circuit is not a perfect inverter; however, it supplies gain, and the two voltages e_{p1} and e_{p2} are both at the same level relative to ground. A disadvantage of the circuit is that it is essentially a feedback circuit, and under certain conditions, particularly when neon-tube couplers are used (see Sec. 2.17), it may oscillate at a high frequency. This can often be corrected, at the expense of poorer frequency response, by connecting a small capacitor between the second grid and ground. It can be shown that the inverting action can be made perfect, i.e., $-e_{p2}/e_{p1}$ can be made exactly unity, by making R_2 slightly greater than R_1 . The precise relation is

$$\frac{R_2 + R_1}{R_2 - R_1} = \frac{\mu R_L}{R_L + r_p} \quad (2.57)$$

2.16. The Cathode-coupled Inverter. A relatively simple inverter, having most of the advantages and only few of the disadvantages of the two inverter circuits discussed thus far, may be obtained by modification of the difference amplifier described in Sec. 2.11. The modification consists of inserting equal plate-load resistors in both plate circuits (Fig. 2.18) and grounding one of the grids. (In practice the grid shown grounded may be connected to a variable-bias voltage and may serve as a balance adjustment.) If the two tubes shown are the two sections of a double triode, it is a good assumption that r_p and μ for the two sections are the same. Note that theoretically the two sections have the same operating point in this circuit.

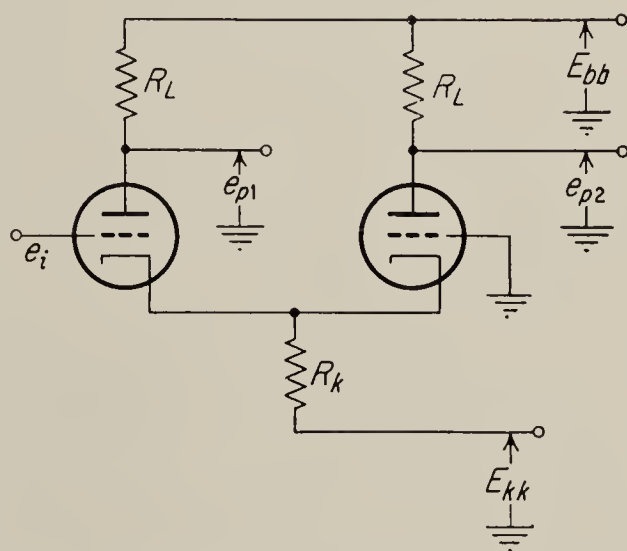


FIG. 2.18. Cathode-coupled inverter.

The equations derived for the difference amplifier apply here, except that r_{p2} must be replaced by $r_p + R_L$. If we make this substitution in Eq. (2.36) and let $e_{g2} = 0$, we have

$$e_{p1} = \frac{-\mu R_L e_i}{R_L + r_p + (\mu + 1)R_k(r_p + R_L)/[(\mu + 1)R_k + r_p + R_L]} \quad (2.58)$$

In computing e_{p2} , e_i becomes the e_{g2} of Eq. (2.36) and e_{g1} is zero; hence

$$e_{p2} = \frac{\mu R_L e_i (\mu + 1) R_k / [(\mu + 1) R_k + r_p + R_L]}{R_L + r_p + (\mu + 1) R_k (r_p + R_L) / [(\mu + 1) R_k + r_p + R_L]} \quad (2.59)$$

Hence
$$\frac{e_{p2}}{e_{p1}} = - \frac{(\mu + 1) R_k}{(\mu + 1) R_k + r_p + R_L} \quad (2.60)$$

Thus this circuit is also not a perfect inverter. However, an increase of R_k will improve the inversion characteristics, and in practice it is found that this circuit inverts at least as well as the paraphase circuit. The circuit has the further advantages of a balanced output, high gain, simplicity, no tendency to oscillate, and the possibility of very simple balance adjustment by use of the second grid.

Where extremely accurate inverting characteristics are required, it is possible to make R_k effectively very large by using a triode as the cathode

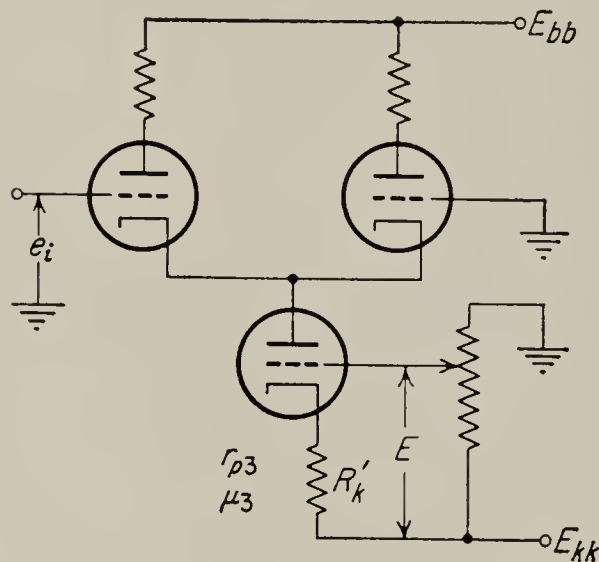


FIG. 2.19. Cathode-coupled inverter with constant-current tube in cathode return.

resistor, as in Fig. 2.19. A pentode would be even better, but the difficulty of supplying its screen voltage and the fact that a triode gives almost perfect inversion usually make the triode more desirable.¹ The effect of this addition to the circuit is simply to replace the R_k in Eq. (2.60) by $r_{p3} + (\mu_3 + 1) R_k$, the resistance seen from the plate of the third tube. Thus this tube results in a very large effective cathode resistor without the disadvantage of the very large negative supply voltage that would be necessary to obtain this same value of R_k without the tube.

2.17. Interstage Coupling Networks. In addition to the problem of drift, the other major problem that distinguishes d-c amplifiers from a-c amplifiers is the design of interstage coupling circuits. These circuits must pass the signal with minimum attenuation and at the same time permit each stage to operate at the correct quiescent level. When the cathodes of all stages are grounded, the requirement is essentially to reduce the quiescent level of the plate of the first stage to that of the grid of the second stage without undue loss of signal.

¹ Another constant-current device that could be used here is the temperature-limited diode. This is a diode whose filament is operated at reduced voltage, so that the plate current is determined by cathode emission and is practically independent of plate voltage. This method, although theoretically very simple, would require a very closely regulated filament voltage for the temperature-limited diode, since the circuit would otherwise exhibit excessive drift.

The simplest method for doing this is the bias battery (see Fig. 2.20). If the quiescent voltage of the plate is e_b volts and that of the grid e_c volts, the battery voltage must be $e_b - e_c$ volts. Ideally the battery has zero internal impedance and perfect voltage stability; under these conditions the coupling is close to ideal and has the following advantages:

1. No signal loss.
2. No power-supply drain.
3. Low-impedance coupling, hence good frequency response and small noise pickup.
4. Coupling introduces no time lag or frequency distortion.

In practice this coupling has several serious disadvantages, and it is used only very rarely. These disadvantages are:

1. Batteries have limited life.
2. Voltage does not remain constant, results in drift.
3. Practical batteries are physically large; hence the capacitance to ground is large, and frequency response is adversely affected. Bulkiness is also a very important disadvantage when miniature size is a requirement.

Another coupling that can be used is the gas-tube coupling (Fig. 2.21). In a voltage amplifier only very little power will be handled by the coupling, and for this reason the smallest neon tubes, such as the NE-2, are used. These tubes have a voltage drop that is almost completely independent of the current through the tube, as long as the region of normal glow is not exceeded.¹ The voltage drop varies considerably from tube to tube, ranging

from about 49 to 74 volts with an average value of 62 volts. The current may vary from 0.03 to 0.3 ma, but the design value is 0.1 ma. In the design of this coupling, one must keep in mind that glow tubes require a relatively high voltage to strike the arc. For the NE-2 neon tube this is about 90 to 100 volts. Although the voltage is nominally independent of current, there is a small increase as the current is increased; the effect is as though there were an internal resistance of about 10,000 ohms. The fact that ionization is involved in the conduction processes through the tube gives rise to a very pronounced time lag

¹ Reich, "Theory and Applications of Electron Tubes," McGraw-Hill Book Company, Inc., New York, 1944, p. 417.

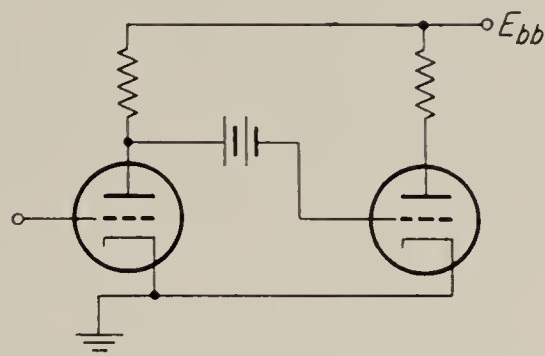


FIG. 2.20. Interstage coupling with bias battery.

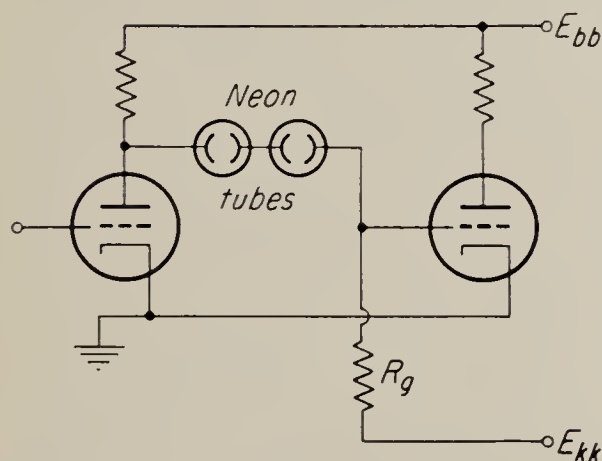


FIG. 2.21. Gas-tube interstage coupling network.

between applied voltage and resulting current at frequencies of about 4,000 cps and higher. This effect may be represented approximately by an inductance whose value depends on the parameters of the rest of the circuit but may reach values of several henrys.

The coupling has the following advantages:

1. Full gain
2. Low-impedance coupling
3. Low power-supply drain
4. Possibility of using only a small negative supply voltage

There are also a number of disadvantages, however, some of them, such as 1, 2, and 3 in the following list, sufficiently serious to outweigh the advantages in many critical applications. These are disadvantages:

1. General unreliability of arc tubes.
2. Small instability of glow results in a certain amount of internally generated noise.
3. Occasionally the circuit breaks into spontaneous relaxation oscillations.
4. Poor high-frequency response due to large inductive effect referred to above.
5. If negative supply voltage is not large enough, glow tubes may sometimes go out during very large signal swings, thus causing very heavy distortion.

In recent years a new and highly efficient coupling device has been developed in the form of the so-called Zener diode. This is a silicon-junction rectifier operated in the reverse direction. Such rectifiers show a very sharp inverse breakdown characteristic in which the voltage is almost exactly constant for large variations of current. Zener diodes are available in voltage ratings ranging from 2 to over 600 volts and with current ratings of 10 μ a to 10 ma. Usually the higher-voltage units must be operated with smaller currents to avoid excessive power dissipation in the rectifier junction. Zener diodes are much more reliable than gas tubes and have a much smaller effective resistance. The resistance is a function of the voltage for which the diode is designed, typical values being about 5 ohms for a 6-volt diode, and 1,000 ohms for a 60-volt diode. Zener diodes generate a certain amount of noise when operated near the knee of their characteristic; however, if sufficient current is permitted to flow through them (typically about 0.2 ma in small diodes) their noise can be reduced to a negligible value. Another disadvantage is that the breakdown voltage is somewhat temperature-dependent. The temperature coefficient for diodes having approximately a 5-volt rating is zero, so that it is theoretically possible to obtain a coupling with zero temperature dependence simply by connecting a sufficient number of 5-volt diodes in series. However, the temperature coefficient of higher-voltage units is

so small that in most applications this refinement is not necessary. Zener diodes are obtainable from most of the manufacturers of semiconductor devices, such as Texas Industries, Transitron, etc. Many of these manufacturers also publish application notes in which the characteristics of the diodes are described in considerable detail.

A very commonly used coupling is the resistance coupling shown in Fig. 2.22. With this coupling there is a signal loss which becomes very marked if E_{kk} is not large. The coupling requires a considerable current, which has the result of decreasing the effective supply voltage to the plate of the first stage (see Sec. 2.25). If the impedance level of R_1 and R_2 is raised to minimize the current drain, a high-impedance coupling results; this is more susceptible to noise pickup. In conjunction with the input capacitance of the following stage, this resistance also represents a low-pass RC filter with a relatively small passband. This results in poor high-frequency response, and although it is possible to compensate for this by bridging R_1 with a capacitor, an optimum design must be made. This is usually a trial-and-error procedure. Despite these disadvantages, this coupling is very commonly used, because it is much more rugged and reliable than the first two described.

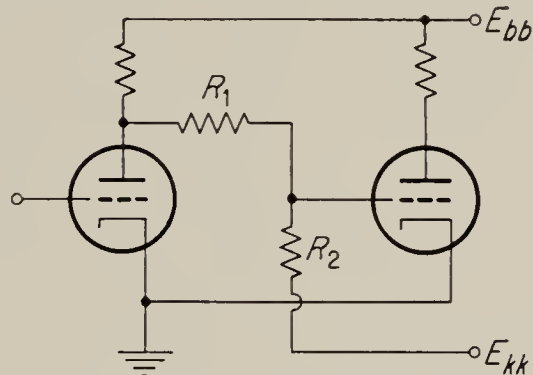


FIG. 2.22. Resistance inter-stage coupling.

If additional tubes are no disadvantage, the *level changer*¹ is occasionally convenient. One form of this circuit, in which a triode is used, is shown in Fig. 2.23. The level changer consists of the tube V_3 and the resistors R_1 , R_2 , R_3 , and R_k . The tube acts essentially as a constant-current device; hence the IR drop through R_1 is constant, and no signal loss occurs. Adjustment of the level may be made by adjusting the grid tap of the level changer. R_1 must be large to prevent excessive current drain, and the tube, as a constant current source, has ideally an infinite resistance [actually $r_p + (\mu + 1)R_k$]. Therefore this coupling is a high-impedance coupling. Hence, except for the fact that the signal is transmitted with no loss, this coupling has the same disadvantages as the resistance coupling.

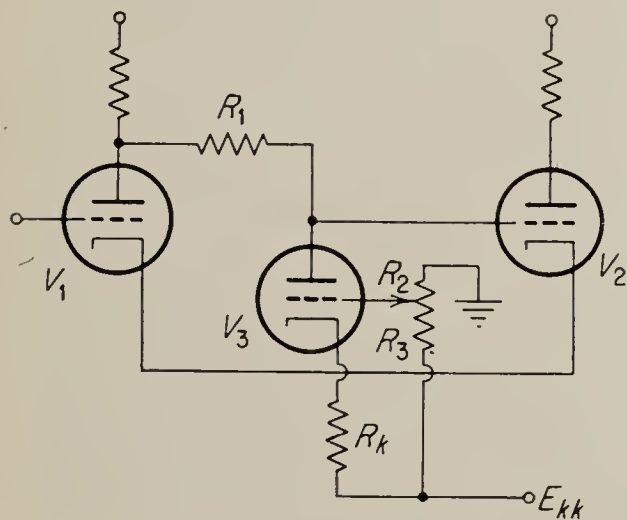


FIG. 2.23. Triode level changer.

Up to now we have considered only circuits in which all the cathodes

¹ See Chestnut and Mayer, "Servomechanisms and Regulating System Design," John Wiley & Sons, Inc., New York, 1955, vol. II, pp. 172-173.

are grounded. However, it is sometimes possible, particularly in push-pull circuits, to permit the cathode voltages of successive stages to become more and more positive, so that it becomes possible to connect the plate directly to the grid of the following stage. This represents an ideal coupling, since it does not attenuate the signal and introduces no impedance. An amplifier embodying this principle is shown in Fig. 2.24.

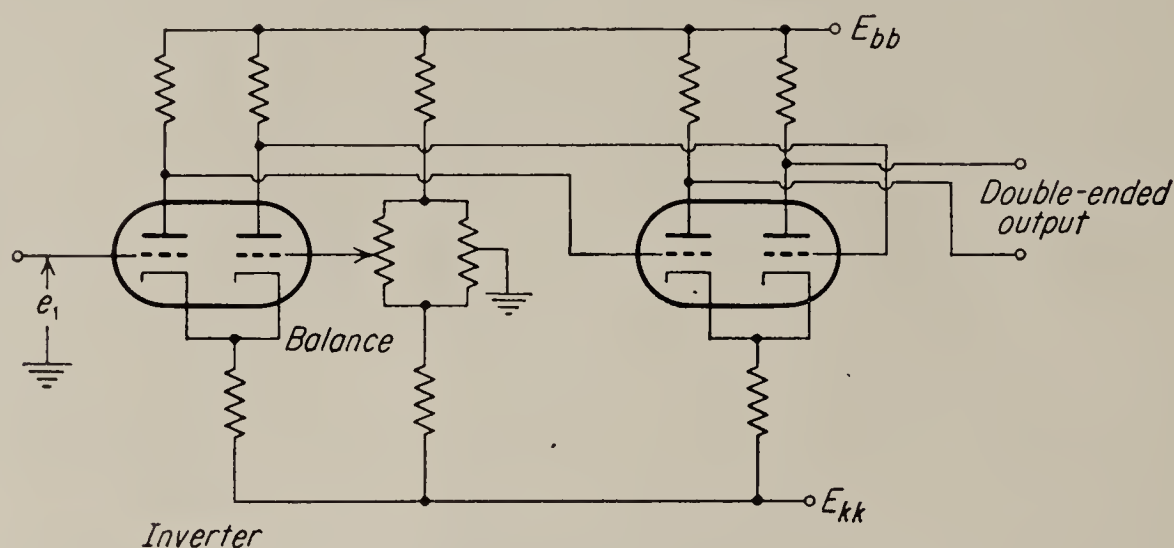


FIG. 2.24. Direct-connected d-c amplifier.

2.18. The Transistor D-C Amplifier. For many control applications transistor d-c amplifiers are very much superior to vacuum-tube amplifiers. Some of the obvious advantages of transistors are their small size and their low power requirements. Since they have no filaments, the large amount of power wasted in heating vacuum-tube filaments may be saved in a transistor amplifier. This reduces the problem of cooling, which becomes quite formidable when many vacuum tubes are crowded into a small space. Also, the absence of filament power removes one source of noise, the 60-cps hum picked up from the filament leads in vacuum-tube circuits. An additional power saving comes from the fact that low-level transistor-amplifier stages operate with collector voltages of a few volts whereas typical plate voltages used on vacuum tubes are on the order of several hundred volts. Transistor amplifiers are potentially much more reliable than vacuum-tube amplifiers. The life of a well-designed transistor is well above 50,000 hours as compared with the 2,000 hours that is standard with vacuum tubes.

Transistors also have, of course, some disadvantages compared to vacuum tubes. The major disadvantage is their sensitivity to temperature variations and the fact that they cannot be used at all at very high temperatures. For germanium transistors the maximum operating temperature is usually well below 100°C , and even for silicon transistors operation at temperatures much above 150°C is usually undesirable. This disadvantage severely limits the application of transistors to control systems. A less important disadvantage is that transistors are electrically

less rugged than vacuum tubes; i.e., they are easily damaged or destroyed by relatively minor overloads. A further disadvantage is the large variation of parameter values among units having nominally the same characteristics. This problem will undoubtedly become less severe as manufacturing and selection processes are improved. Transistors cannot be used at very high radio frequencies but this is of no consequence in a control system. Finally, transistors are extremely sensitive to nuclear radiation. Since transistor action takes place by means of minute atomic impurities (the positive and negative charge carriers referred to below), any disturbance of this action by radiation particles has a significant effect on the operation of the device.

A junction transistor may be thought of as a sandwich of three sections of semiconductor crystal. The central section is referred to as the *base* while the outside sections are the *emitter* and *collector* respectively. In most transistors the junction between collector and base has a larger area than that between emitter and base, but except for this the junctions are identical. There are two types of transistors. In the *p-n-p* type the collector and emitter consist of material in which the electric current is carried by *holes*, or positive charge carriers, while in the base the current is carried by *electrons* or negative carriers. The other transistor type is the *n-p-n*, where the base consists of positive material and the emitter and collector of negative material.

The operation of a transistor may be explained by thinking of the two junctions as rectifying junctions. The collector junction is biased in the reverse direction, and in the absence of emitter current only a very small inverse current flows in the collector circuit. The emitter junction is biased in the forward direction, so that only a small voltage between emitter and base is needed to cause relatively large emitter currents to flow. The flow of emitter current through the base results in the production of a large number of charge carriers within the base, and these are picked up by the collector, resulting in a flow of current in the collector. Thus, we may define *transistor action* as the control of current in the negatively biased collector junction by current in the positively biased emitter junction. (The terms "positive" and "negative" are correct for *p-n-p* transistors; they would be reversed for *n-p-n* transistors.) An important transistor parameter is α , the ratio of collector current to emitter current. In junction transistors α is slightly less than one, typical values ranging from 0.95 to 0.98. In point-contact transistors α normally exceeds one.

2.19. Analysis of Simple Transistor Circuits. In analyzing transistor circuits it is convenient to assume small signal operation about a fixed operating point just as in vacuum tubes. In the following discussion we consider only junction transistors, since point-contact transistors are not

used in control amplifiers. Also, the transistors will be assumed to be of the p - n - p variety unless a statement to the contrary is made.

Consider a transistor connected, as shown in Fig. 2.25a, with the base grounded. This circuit may be analyzed as a standard four-terminal network if we assume that the emitter, base, and collector all have an

electrode resistance r_e , r_b , and r_c respectively, and that in addition, the collector voltage v_2 for fixed collector current i_c is given by the product of the emitter current i_e and a fictitious resistance which, for the moment, will be designated by r_x . We obtain

$$v_1 = (r_e + r_b)i_e + r_b i_c \quad (2.61)$$

$$v_2 = r_x i_e + (r_b + r_c)i_c \quad (2.62)$$

It should be noted that none of the four resistances, r_e , r_b , r_c , or r_x , have any physical existence within the transistor; they are in fact defined by Eqs. (2.61) and (2.62). Explicitly the definitions are

$$r_b \triangleq \left. \frac{\partial v_1}{\partial i_c} \right|_{i_e = \text{const.}} \quad (2.63)$$

$$r_e \triangleq \left. \frac{\partial v_1}{\partial i_e} \right|_{i_c = \text{const.}} - r_b \quad (2.64)$$

and

$$r_c \triangleq \left. \frac{\partial v_2}{\partial i_c} \right|_{i_e = \text{const.}} - r_b \quad (2.65)$$

FIG. 2.25. Transistor in grounded-base connection: (a) actual circuit; (b) a-c equivalent circuit; (c) d-c equivalent circuit showing I_{co} .

In the definition of r_x we prefer to introduce the current gain α already mentioned above. For the directions assumed for the currents i_c and i_e , this parameter is defined by

$$\alpha = - \left. \frac{\partial i_c}{\partial i_e} \right|_{v_2 = \text{const.}} \quad (2.66)$$

since the actual collector current in a p - n - p transistor is out of the collector. From Eq. (2.62) we obtain that

$$\left. \frac{\partial i_c}{\partial i_e} \right|_{v_2 = \text{const.}} = - \frac{r_x}{r_b + r_c} \quad (2.67)$$

Hence we have that

$$\frac{r_x}{r_b + r_c} = \alpha$$

or

$$r_x = \alpha(r_b + r_c) \quad (2.68)$$

A possible equivalent circuit from which Eqs. (2.61) and (2.62) could have been obtained directly is shown in Fig. 2.25b. This equivalent circuit is not quite correct for d-c amplifiers since it indicates that $i_c = 0$ when $i_e = 0$. In practice a small current I_{co} flows in the collector circuit even when $i_e = 0$. To account for this current the circuit of Fig. 2.25b may be modified as shown in Fig. 2.25c.¹ The magnitude of I_{co} depends on the transistor and the temperature. In a typical transistor (General Electric 2N43A) I_{co} ranges between 5 and 10 μ a at a temperature of

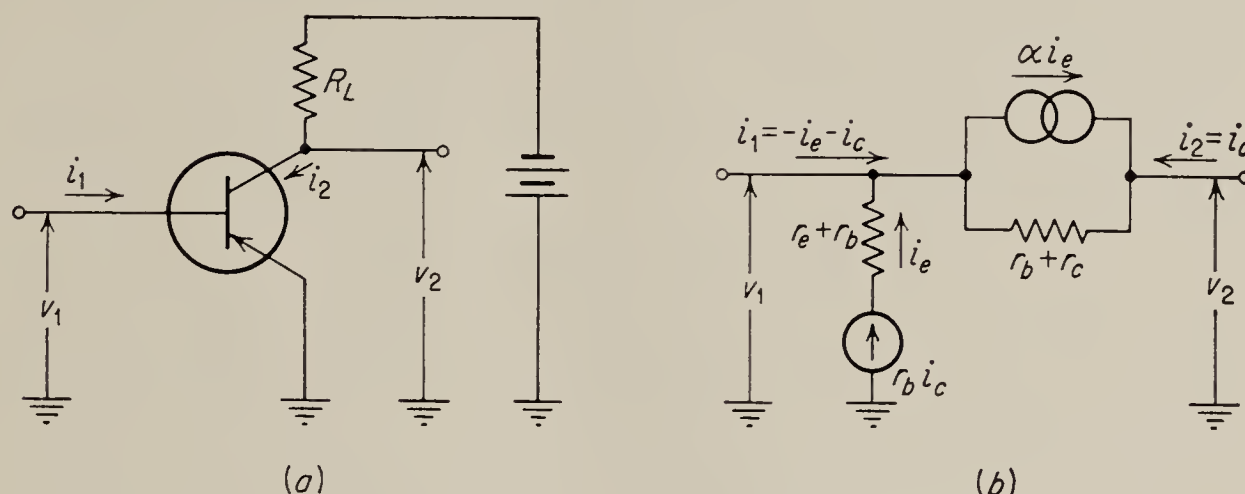


FIG. 2.26. The grounded-emitter circuit: (a) actual circuit showing load resistance and battery; (b) equivalent circuit.

25°. The temperature dependence of I_{co} is one of the major factors causing transistor d-c amplifiers to drift. The variation of I_{co} is given approximately by

$$I_{co} = I_{coo} e^{k(T-T_o)} \quad (2.69)$$

where $I_{coo} = I_{co}$ at the temperature T_o , and k varies from 0.06 to 0.09 with an average value being 0.08.* This relation is valid for T_o of about 25°C and $T - T_o$ less than 50°C. It should be noted that for $k = 0.07$, I_{co} doubles for every 10°C change in temperature.

Since the grounded-base connection discussed up to now has a current gain of less than one, direct coupling of stages cannot produce current gains greater than one, and the voltage or power gain for a multistage amplifier would therefore be the same as for a single stage. For this reason the most commonly used connection is the grounded-emitter connection shown in Fig. 2.26a. This connection is closely analogous to the grounded-cathode connection for vacuum tubes. To obtain the equations for this connection we may redraw the equivalent circuit of Fig. 2.25b in the form shown in Fig. 2.26b. Note that the input current is now $i_b = i_1 = -i_e - i_c$, and the output current is $i_c = i_2$ as before.

While it would be possible to write the equations for this circuit in the same form as Eqs. (2.61) and (2.62), i.e., with the voltages as the depend-

¹ Hunter, "Handbook of Semiconductor Electronics," McGraw-Hill Book Company, Inc., New York, 1956, pp. 13.2-13.4.

* *Ibid.*

ent variables, it is desirable to write them in a different form in which the input voltage and output current are the dependent variables:

$$v_1 = h_{ie}i_1 + h_{re}v_2 \quad (2.70)$$

$$i_2 = h_{fe}i_1 + h_{oe}v_2 \quad (2.71)$$

The h parameters appearing in these equations can be defined in terms of r_e , r_b , r_c , and α used above.¹ A possible procedure is to first set up the equations for the equivalent circuit of Fig. 2.26b with v_1 and v_2 as the dependent variables:

$$v_1 = (r_e + r_b)i_1 + r_e i_2$$

$$v_2 = [r_e + r_b - \alpha(r_b + r_c)]i_1 + [r_c + r_b + r_e - \alpha(r_b + r_c)]i_2$$

Rearranging gives

$$-(r_e + r_b)i_1 = -v_1 + r_e i_2$$

$$v_2 - [r_e + r_b - \alpha(r_b + r_c)]i_1 = [r_c + r_b + r_e - \alpha(r_b + r_c)]i_2$$

Solving for v_1 and i_2 results in Eqs. (2.70) and (2.71), and by direct comparison we find that

$$\begin{aligned} h_{ie} &= r_e + r_b - \frac{r_e[r_e + r_b - \alpha(r_b + r_c)]}{r_c + r_b + r_e - \alpha(r_b + r_c)} \approx \frac{r_e}{1 - \alpha} + r_b \\ h_{re} &= \frac{r_e}{r_c + r_b + r_e - \alpha(r_b + r_c)} \approx \frac{r_e}{r_c(1 - \alpha)} \\ h_{fe} &= -\frac{r_e + r_b - \alpha(r_b + r_c)}{r_c + r_b + r_e - \alpha(r_b + r_c)} \approx \frac{\alpha}{1 - \alpha} \\ h_{oe} &= \frac{1}{r_e + r_b + r_c - \alpha(r_b + r_c)} \approx \frac{1}{r_c(1 - \alpha)} \end{aligned} \quad (2.72)$$

In making the approximations, advantage has been taken of the fact that r_c is usually very large, in the megohm range, while r_b and r_e are typically less than 1,000 and 50 ohms respectively.

The equations for the transistor in terms of the h parameters and the subscript notation used are now being accepted as standard by industry. The subscripts i , r , f , and o refer to input, reverse, forward, and output respectively, while the subscript e refers to the fact that the parameters are defined for the grounded-emitter connection. For the grounded-base and grounded-collector connections the subscripts b and c are used. A double subscript notation using the numbers 1 and 2, i.e., h_{11} , h_{12} , h_{21} , h_{22} , instead of h_i , h_r , h_f , h_o , is also in common use, and for some applications z and y parameters are occasionally used.²

¹ The h parameter representation is one of the standard ways of representing a four-terminal network. Other methods include z parameters, y parameters, etc. See Guillemin, "Communications Networks," John Wiley & Sons, Inc., New York, 1935, pp. 134-138.

² Hunter, *op. cit.*, sec. 11. Also see Shea, "Transistor Circuit Engineering," John Wiley & Sons, Inc., New York, 1957, p. 22.

In terms of the h parameters, the equivalent circuit of the transistor takes the form of Fig. 2.27a. Typical values for the h parameters are also given in this figure.¹ Note that h_{ie} is a resistance while h_{oe} is a conductance. The advantage of the h parameter representation is that in most practical circuits the load admittance $1/R_L$, which effectively shunts h_{oe} (see Fig. 2.26a), is much greater than h_{oe} , so that h_{oe} can often be

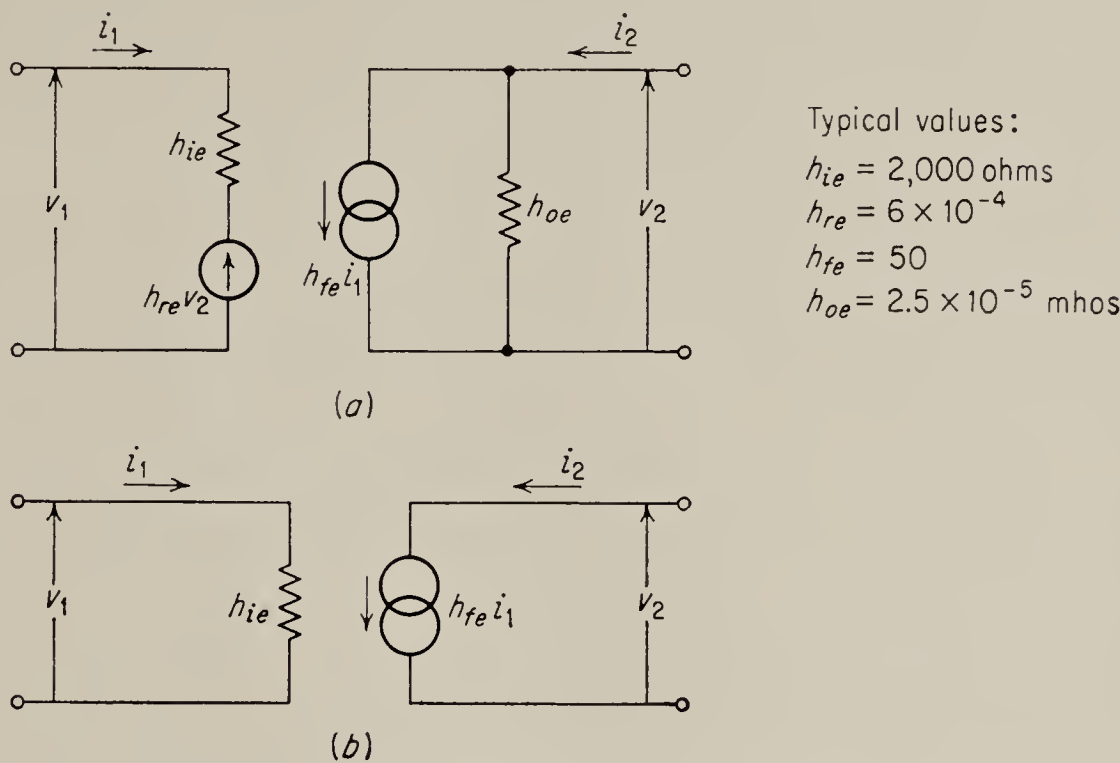


FIG. 2.27. Equivalent circuit for grounded-emitter circuit in terms of h parameters: (a) exact circuit; (b) approximate circuit.

neglected. Similarly the voltage $h_{re}v_2$ in the input circuit is usually much less than the voltage drop through h_{ie} , so that the input impedance is approximately h_{ie} . Thus the very simple approximate equivalent circuit of Fig. 2.27b results. In this circuit the input circuit is completely decoupled from the output, making the transistor a unilateral current amplifier with the current gain h_{fe} .

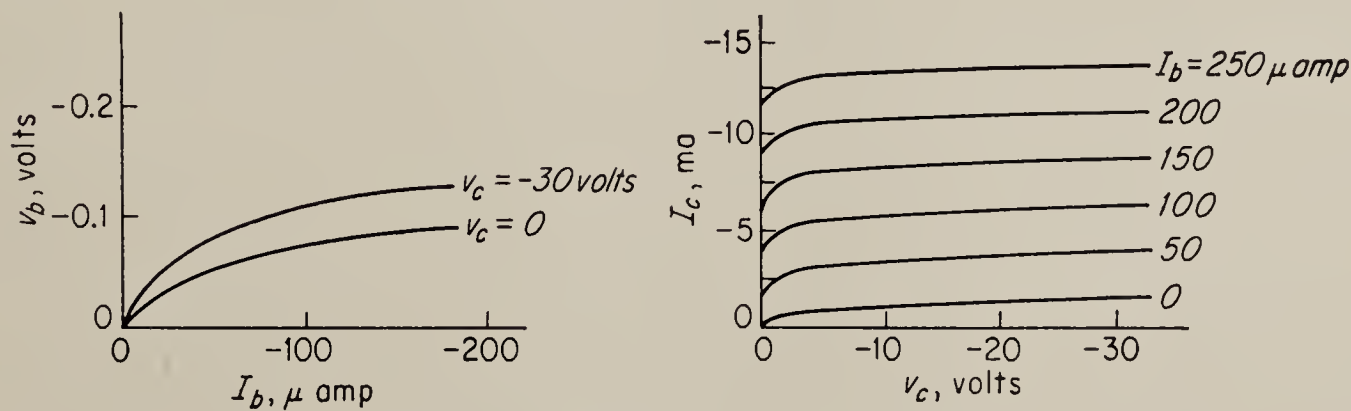


FIG. 2.28. Characteristic curves for grounded-emitter connection.

Most manufacturers now supply values for the h parameters of the transistors they produce rather than values for r_e , r_b , or r_c . Also, the h parameters can be read directly from the characteristic curves usually furnished. A typical set of characteristics is shown in Fig. 2.28. Note

¹ Hunter, *op. cit.*, p. 11.6.

that two families of curves are required: an input family and an output family. From the input family one obtains h_{ie} , the slope of the curve at the chosen operating point, and h_{re} , the vertical distance between two curves for which the difference in collector voltage is unity. Similarly, from the output family one obtains h_{fe} , the vertical distance between curves for which the difference in base current is unity, and h_{oe} , the slope of the curve. From the shape of the curves it is apparent that the parameters h_{ie} and h_{fe} can be obtained with good accuracy, while h_{re} and h_{oe} are difficult to get accurately. Fortunately, the important parameters are the ones that can be obtained accurately. It should also be noted that the output characteristics are almost parallel and equidistant for constant increments of base current. This means that the transistor is a good linear amplifier of the input current; i.e., a sinusoidal input current will result in sinusoidal output current or voltage. On the other hand, due to the curvature of the input characteristic, a sinusoidal voltage applied to a transistor will in general yield a nonsinusoidal output. It is therefore usually desirable to drive the base from a high-impedance, or constant-current, source. It should be pointed out that the h parameters are basically small-signal parameters just like the μ , r_p , and g_m of a vacuum tube, and that they vary with the operating point. They also vary with the operating temperature. Thus the typical values given in Fig. 2.27 can vary by factors of two or three or more. Most manufacturers of transistors do, however, supply charts showing what this variation is.

The equivalent circuit of Fig. 2.27a again does not take into account the fact that with zero base current the collector current is not zero. In order to obtain an equivalent circuit that properly takes into account all of the d-c conditions of a transistor, we may go back to the circuit of Fig. 2.25c, and convert it to the common-emitter form in terms of h parameters. The details of this conversion are left to the reader, but it results approximately in introducing an additional voltage source $-I_{co}h_{fe}h_{re}/h_{oe}$ into the input circuit, and an additional current source $-h_{fe}I_{co}$ into the output circuit (see Prob. 2.9).

2.20. Single-stage Circuits. The current gain, voltage gain, power

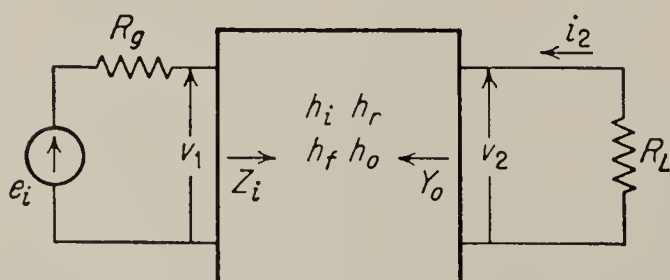


FIG. 2.29. Four-terminal network.

gain, input impedance, and output impedance of a transistor may be easily obtained from the four-terminal representation shown in Fig. 2.29. In this representation it is assumed that the currents i_1 and i_2 and the voltages v_1 and v_2 are related by Eqs. (2.70) and (2.71).

In order to find the current gain, we note that in Fig. 2.29, $v_2 = -i_2R_L$. Substituting the value of v_2 in Eq. (2.71), we immediately find that

$$\frac{i_2}{i_1} = G_i = \frac{h_{fe}}{1 + h_{oe}R_L} \approx h_{fe} \quad (2.73)$$

The approximation is justified since typically $R_L h_{oe} \ll 1$. (See typical values of the h parameters in Fig. 2.27a; R_L is usually less than 5,000 ohms.)

To obtain the voltage gain we replace i_2 in Eq. (2.71) by $-v_2/R_L$, and solve Eqs. (2.70) and (2.71) simultaneously. The result is

$$\frac{v_2}{v_1} = G_v = \frac{-h_{fe}(R_L/h_{ie})}{1 + R_L h_{oe} \left(1 - \frac{h_{fe} h_{re}}{h_{oe} h_{ie}}\right)} \quad (2.74)$$

Here again a simple approximate result may be obtained by letting $R_L h_{oe} \ll 1$, and noting that $h_{fe} h_{re}/h_{oe} h_{ie}$ is typically about 0.6. The approximate result is

$$G_v \approx -h_{fe}(R_L/h_{ie}) \quad (2.75)$$

By a similar procedure we may find that the input impedance may be written

$$Z_i = \frac{v_1}{i_1} = h_{ie} \left[1 - \frac{(h_{fe} h_{re})(R_L h_{oe})}{(h_{ie} h_{oe})(1 + R_L h_{oe})} \right] \approx h_{ie} \quad (2.76)$$

and the output admittance is

$$Y_o = \frac{i_2}{v_2} = h_{oe} \left[1 - \frac{(h_{fe} h_{re})(h_{ie})}{(h_{ie} h_{oe})(h_{ie} + R_g)} \right] \approx h_{oe} \quad (2.77)$$

The power gain is simply the product of current and voltage gain. It should be noted that the expressions for current gain, input impedance, and output impedance reduce approximately to one of the h parameters. Also the voltage gain is approximately equal to the product of current gain and ratio of load resistance to input resistance. These approximate results could have been obtained directly by use of the approximate equivalent circuit of Fig. 2.27b.

Although the grounded-emitter circuit is the one used most often, the grounded-collector and grounded-base circuits are also used occasionally. The form of these circuits for p - n - p transistors is shown in Fig. 2.30. For n - p - n transistors, the battery voltage should be reversed. The simplest way to obtain the various gains and impedances for these circuits is first to compute new h parameters for them. These h parameters may then be used instead of the common-emitter parameters in Eqs. (2.73), (2.74), (2.76), and (2.77). This is possible since these equations describe the characteristics of the generalized four-terminal network of Fig. 2.29. It must be noted, however, that the approximations made in these four equations apply in general only to the grounded-emitter connection.

As an example of the process of obtaining the modified h parameters, we consider the common collector circuit of Fig. 2.30b. By comparison with Fig. 2.26a we note that

$$\begin{aligned} v_{1e} &= v_{1c} - v_{2c} \\ v_{2e} &= -v_{2c} \\ i_{1e} &= i_{1c} \\ i_{2e} &= -i_{1c} - i_{2c} \end{aligned}$$

where the subscripts e refer to grounded emitter and c to grounded collector. Rewriting the equations for the grounded-emitter connection, Eqs.

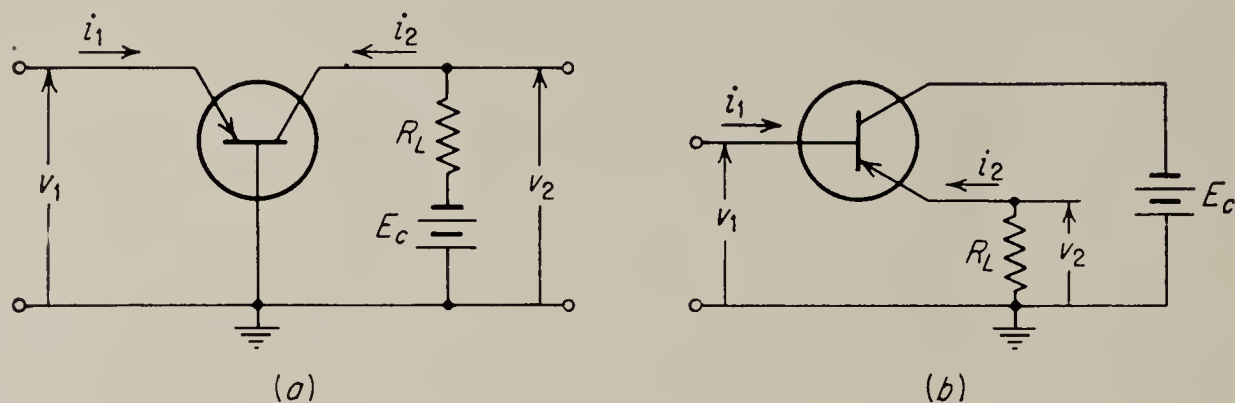


FIG. 2.30. (a) Grounded-base circuit; (b) grounded-collector circuit.

(2.70) and (2.71), and rearranging results then directly in

$$\begin{aligned} v_{1c} &= h_{ie}i_{1c} + (1 - h_{re})v_{2c} \\ i_{2c} &= -(1 + h_{fe})i_{1c} + h_{oe}v_{2c} \end{aligned}$$

by direct comparison, we obtain:

$$\begin{aligned} h_{ic} &= h_{ie} \\ h_{rc} &= 1 - h_{re} \approx 1 \\ h_{fc} &= -(1 + h_{fe}) \\ h_{oc} &= h_{oe} \end{aligned} \tag{2.78}$$

By a similar but somewhat more complicated process, the common-base parameters may be obtained in terms of the common-emitter parameters. For the common-base connection we have approximately

$$\begin{aligned} h_{ib} &= h_{ie}/(1 + h_{fe}) \\ h_{rb} &= \frac{h_{ie}h_{oe}}{1 + h_{fe}} - h_{re} \\ h_{fb} &= -\alpha = -\frac{h_{fe}}{1 + h_{fe}} \\ h_{ob} &= \frac{h_{oe}}{1 + h_{fe}} \end{aligned} \tag{2.79}$$

The approximations used in obtaining the common-base parameters are that $1 \gg h_{re}$ and that $h_{ie}h_{oe} \ll h_{fe}$. Both of these are justified as may be seen from the typical values of the h_e parameters in Fig. 2.27a.

The modified h parameters may now be used in Eqs. (2.73) to (2.77) to find the current and voltage gain, input impedance, and output admittance for the various connections. The exact expressions for these characteristics will have exactly the same appearance for all three connections, if the modified h parameters are used, but the approximations that can be made are quite different for the three connections. In Table 2.1 the

TABLE 2.1

	Z_i	Z_o	G_i	G_v	Approximations made
Common base	$\frac{h_{ie}}{1 + h_{fe}}$	$\frac{1 + h_{fe}}{h_{oe}}$	$-\frac{h_{fe}}{1 + h_{fe}}$	$(1 + h_{fe}) \frac{R_L}{h_{ie}}$	$R_L h_{oe} \ll 1$ $R_g/h_{ie} \gg 1$ $0 < \frac{h_{fe} h_{re}}{h_{ie} h_{oe}} < 1$
Common emitter	h_{ie}	$\frac{1}{h_{oe}}$	h_{fe}	$-h_{fe} \frac{R_L}{h_{ie}}$	$R_L h_{oe} \ll 1$ $R_g/h_{ie} \gg 1$ $0 < \frac{h_{fe} h_{re}}{h_{ie} h_{oe}} < 1$
Common collector	$R_L(1 + h_{fe})$	$\frac{R_g}{1 + h_{fe}}$	$-(1 + h_{fe})$	1	$R_L h_{oe} \ll 1$ $R_g/h_{ie} \gg 1$ $(1 + h_{fe}) R_L/h_{ie} \gg 1$ $(1 + h_{fe})/(R_g h_{oe}) \gg 1$

approximate characteristics for all three connections are summarized. The expressions are all in terms of common-emitter h parameters. The approximations made for each connection are also given. It should be noted that these approximations are not applicable in all circuits of practical importance, particularly the requirement that $R_g/h_{ie} \gg 1$ is not always met. We note that the input impedance is very small (ca. 50 ohms) for the grounded base, intermediate for the grounded emitter, and highest for the grounded collector. The lowest output impedance is found in the common collector, and the highest in the grounded base. Only the common-emitter connection has a voltage and current gain greater than unity; this is the chief reason why it is used more frequently than the other two connections.

It is often necessary, particularly in d-c amplifier circuits, to consider circuits having resistances other than the load resistance R_L . Particularly important is the common-emitter circuit with a resistor in the emitter lead, the common emitter with a feedback resistor between collector and base, and the common collector with a resistor in the collector lead. All of these modifications of the basic circuits are most conveniently handled by computing modified h parameters for them. This is

done in a manner similar to that explained in connection with Eq. (2.78). A complete list of modified h parameters is given by Hunter.¹ In general the effects of additional resistors of this sort are the same as in vacuum-tube circuits. For instance, a resistor in the emitter lead of a common-emitter connection results in a marked decrease in voltage gain, and an increase in the input and output impedance.

2.21. Drift in Transistor Amplifiers. The main cause of drift in transistor amplifiers is the temperature variation of I_{co} , the collector-to-base current that flows when the emitter current is zero. The variation of this current with temperature is given in Eq. (2.69). As was pointed out in connection with this equation, I_{co} approximately doubles for every 10°C change in temperature. A second effect causing drift is the change of the h parameters themselves with temperature. This is quite severe in some transistors. The drift due to I_{co} may result in a condition of thermal runaway where an initial increase of I_{co} due to an external temperature rise results in a further temperature rise in the transistor, which, in turn, results in a further increase in I_{co} , etc., until the transistor either destroys itself or drifts into a relatively inoperative condition. Even if thermal runaway does not occur, the variation in the h parameters and in I_{co} generally tends to shift the operating point of the transistor away from the optimum value, and thus to magnify the drift.

There are several ways of combating drift. An obvious method of alleviating the problem is to operate the transistors with a large enough collector current so that I_{co} is only a small fraction of the total collector current. In this way the percentage change of collector current resulting from a given change in temperature may be reduced. A related method is to design the networks that supply bias to the terminals in such a way that changes in I_{co} result in only small changes of operating point. Shea² defines a stability factor for current as the ratio of a change in the quiescent operating value of the emitter to a temperature-induced variation in I_{co} . Similarly, he defines a voltage stability factor for the collector-to-base voltage. These stability factors can be expressed in terms of the resistors of the biasing network and the power-supply voltages. In designing a biasing network, limits on the permissible shift of operating point are first set by taking into account the permissible changes in the transistor parameters. These limits together with the variation in I_{co} expected (a function of the expected temperature change) are then used to set an upper limit on the stability factors. The biasing network may then be designed by the use of the relationships existing between the stability factors and the components of the biasing network. Complete design tables are given by Shea, and the reader is referred to this book

¹ Hunter, *op. cit.*, pp. 11.26–11.27.

² Shea, *op. cit.*, chap. 3.

for details of the procedure. For the common emitter circuit, Shea's results can be summarized as follows:

1. The poorest situation results when the base is driven from a high-resistance source, and the emitter is grounded.
2. Placing a resistor in the common emitter lead improves stability.
3. Reducing the output resistance of the driving source improves stability.
4. A feedback resistor connected between the collector and base improves stability.

It should also be noted that a large load resistor in the collector lead tends to protect a transistor from thermal runaway, since an increase in I_{co} causes a decrease of collector voltage. This is true even though the voltage stability factor as defined by Shea increases with R_L .

Another method of reducing drift is to use a difference amplifier¹ as shown in Fig. 2.31. This circuit is completely analogous to the vacuum-tube difference amplifier discussed in Sec. 2.11, and its output is effectively equal to the difference in the two base signals multiplied by a gain factor. Also, as far as changes in I_{co} are concerned, the output will be a function of the difference of the currents in the two transistors. Thus the circuit will discriminate against variations in I_{co} in the same way as the tube circuit of Sec. 2.11 discriminates against changes in e_h . The complete analysis of the transistor version of the circuit is left as an exercise for the reader.

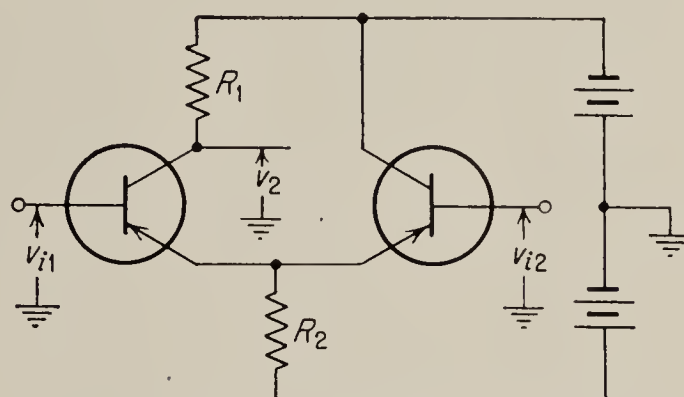


FIG. 2.31. The transistor difference amplifier.

The difference amplifier is only one special case of using a transistor as a compensating element to reduce drift in a d-c amplifier. Other methods of using transistors, diodes, and temperature-sensitive elements in this way are described in the literature.² Most of these methods require that the amplifier be adjusted for least drift after construction.

2.22. Multistage Amplifiers. When transistor amplifier stages are cascaded to form a multistage amplifier it is necessary, just as in vacuum-tube amplifiers, to design interstage coupling networks which will permit proper voltage and current biases to be applied to the coupled transistors. This problem is, however, often much simpler in transistor amplifiers than in vacuum-tube amplifiers. This is due to three factors. First the col-

¹ D. W. Slaughter, Feedback Stabilized Transistor Amplifier, *Electronics*, vol. 28, May, 1955, pp. 174-175.

² Shea, *op. cit.*, chap. 5. Hunter, *op. cit.*, p. 13.17. E. R. Kretzmer, An Amplitude Stabilized Transistor Oscillator, *Proc. IRE*, vol. 42, pp. 391-401, 1954.

lector may often be operated at voltages of only 1 or 2 volts, particularly in low-level stages. Secondly, if transistors of the same type (either all $p-n-p$ or all $n-p-n$) are used, the base and collector voltages have the same polarity. Finally, it is possible to alternate between $p-n-p$ and $n-p-n$ transistors.

A typical three-stage circuit illustrating some of these points is shown in Fig. 2.32. In the first stage a resistor is employed in the common-emitter lead to provide some bias stabilization. In picking a value for this resistor it should be kept in mind that a very large value will increase the input impedance and reduce the voltage gain, but that it does not have much effect on the current gain. As long as the input impedance of the transistor is less than the source resistance of the generator R_g , the

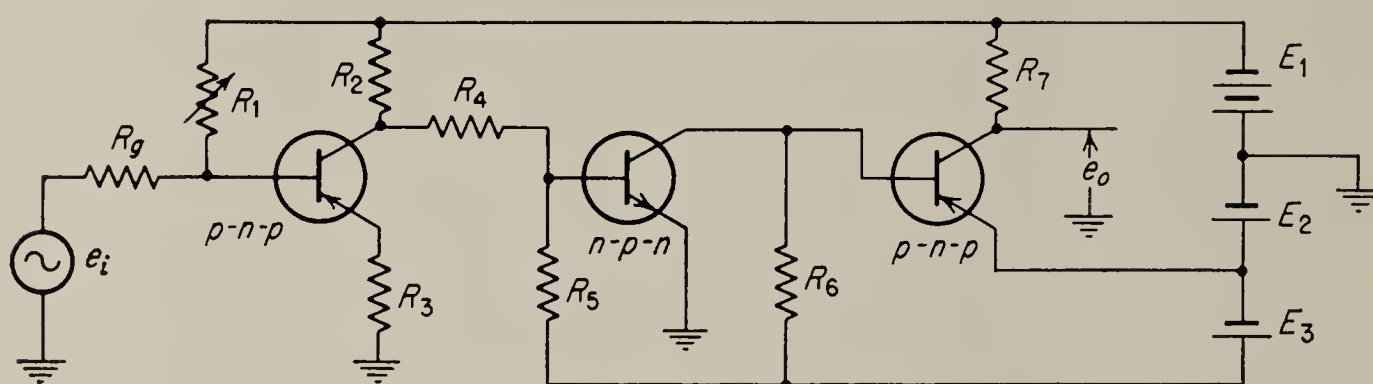


FIG. 2.32. Typical transistor three-stage d-c amplifier.

current delivered by the generator and the current amplification of the stage will not be reduced by R_3 . Thus a compromise value that will not make the input impedance too high should be used. Typically R_3 is on the order of 1,000 ohms. The plate-load resistor R_2 may be chosen in the same way as in vacuum-tube circuits, by use of the load line. It should not be too large, so that the quiescent collector current is very much greater than the current flowing for zero base current. The biasing resistor R_1 is now chosen to produce the correct emitter current as demanded by the choice of operating point. It may be computed by assuming that the base-to-emitter voltage is approximately zero. For zero e_i the quiescent value of the emitter current I_e is then equal to

$$I_e = \frac{E_1}{R_1 + (R_g R_3)/(R_g + R_3)} \frac{R_g}{R_g + R_3} = \frac{E_1 R_g}{R_1(R_g + R_3) + R_g R_3}$$

Since E_1 , I_e , R_g , and R_3 are known, R_1 may be found. The operating point of the first stage, and therefore the d-c level of the amplifier output, can be varied by adjusting R_1 .

The second stage is an $n-p-n$ stage; hence the base must be biased in the positive direction. If the positive voltage $E_3 + E_2$ is very much greater in size than the negative collector voltage of the first stage, the gain loss in the coupling network is small. This fact would dictate a choice of small collector operating voltage for the first stage. R_4 and R_5 are

designed so that the base of the second stage operates at the correct value of quiescent emitter current.

In the coupling between the second and third stages no resistance is needed at all since the collector of the second stage and the base of the third stage can both operate at the same positive voltage (approximately E_2). As a matter of fact, in low-level stages it is sometimes possible to dispense also with resistor R_6 and to let the input impedance of the third stage act as the collector load resistance for the second stage. A very simple circuit results.

The emitter of the third stage is connected to the positive voltage E_2 so that the output voltage e_o can be zero for zero e_i . Instead of using a separate battery E_2 , a bleeder may be used between the positive supply and ground, or more elegantly, a silicon diode (Zener diode) may be used instead.¹

2.23. The Chopper-stabilized D-C Amplifier. Although the difference amplifier and the Miller circuit discussed in previous sections are capable

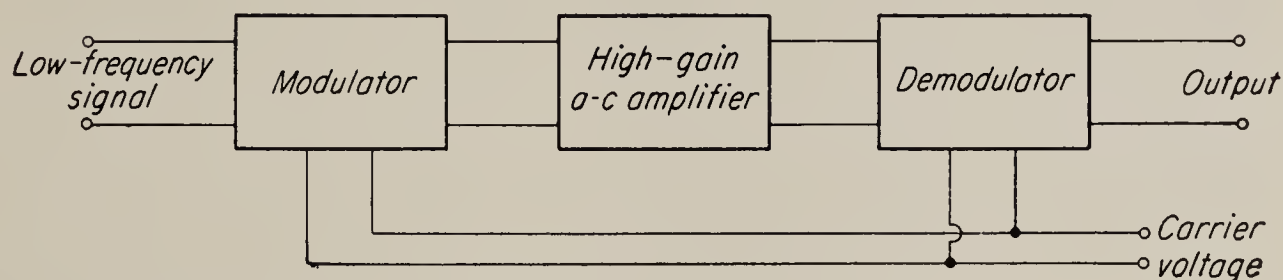


FIG. 2.33. D-c amplifier using modulation and demodulation.

of reducing drift to a value low enough for d-c amplifiers of moderate gain, these relatively simple circuits cannot meet the requirement for almost zero drift that is made in the most critical situations. D-c amplifiers used in analogue computers or amplifiers used to amplify the extremely minute voltages generated by thermocouples are representative of this class of problem.

For such applications it is necessary to “detour” around the drift problem introduced by d-c amplifiers and to use a-c amplifiers instead. A typical circuit to accomplish this is shown in block-diagram form in Fig. 2.33. Here the signal is passed first into a modulator, which converts the d-c (or very low frequency) signal into a relatively high frequency signal that may then be amplified in a standard a-c amplifier. Since a-c amplifiers with very high gain are relatively easy to construct and since they do not drift, a very large amount of drift-free amplification may be obtained, provided that the modulator itself is free of drift. After the signal has been amplified to the desired level by the a-c amplifier, it is passed through a demodulator to recover the original d-c, or low-frequency, information.

Although a number of electronic modulator and demodulator circuits

¹ Hunter, *op. cit.*, p. 13.10.

are available (see Chap. 6), most of them are subject to some drift. Hence, for critical applications the *electromechanical vibrator*, or *chopper*, is used (see Fig. 2.34). These devices have in recent years achieved a very high degree of perfection; they are hermetically sealed and sufficiently rugged to withstand large amounts of acceleration, shock, and

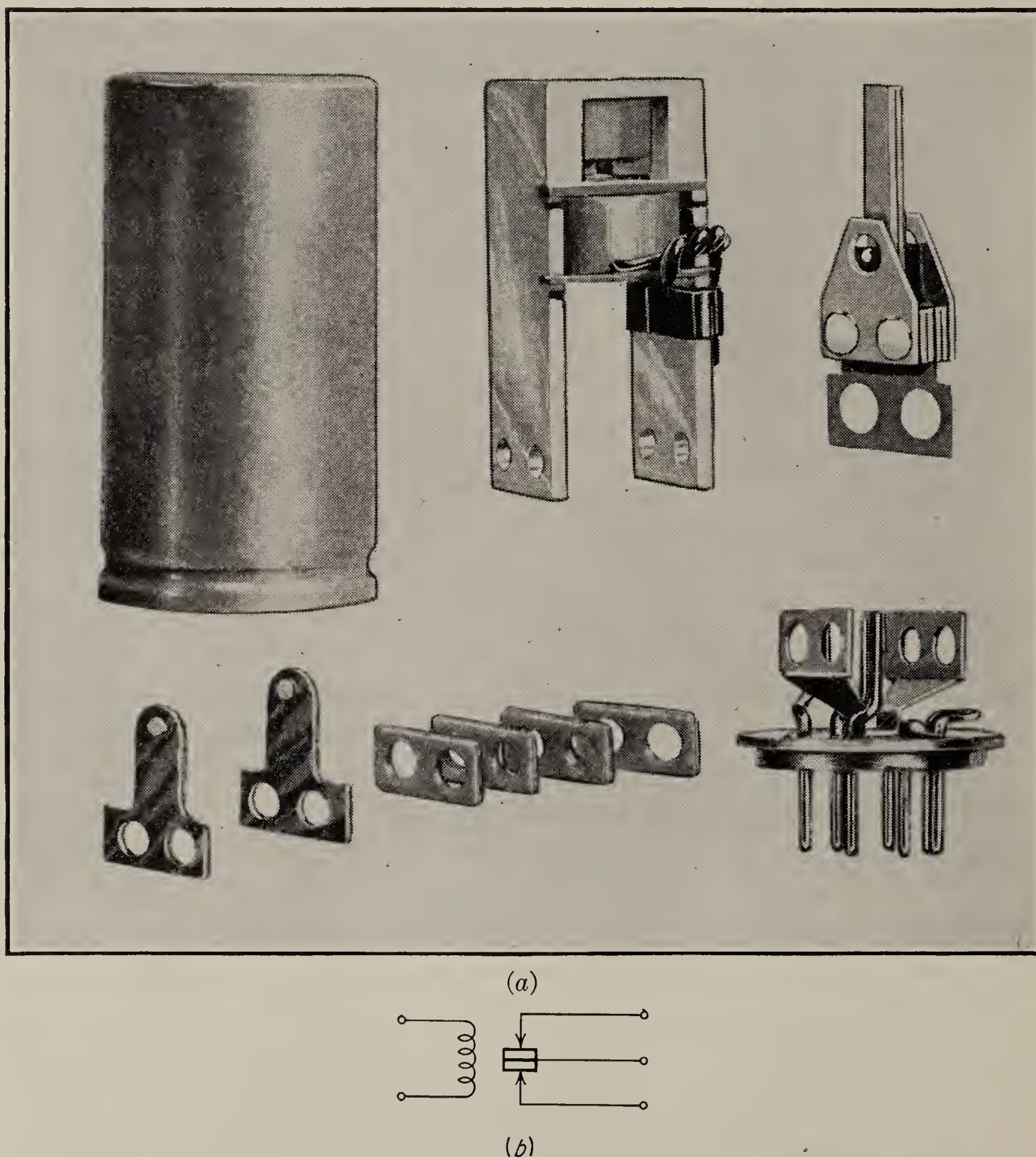


FIG. 2.34. Electromechanical chopper: (a) disassembled; (b) schematic.

temperature variation. The terminals are usually brought out to a standard tube socket, and the outside dimensions are comparable to those of standard or even miniature vacuum tubes. Further, their life expectancy under normal operating conditions is comparable to that of vacuum tubes, and they may, therefore, be designed into a circuit in exactly the same way as a tube would be.

A convenient method of connecting the chopper is shown in Fig. 2.35, where a single chopper performs the functions of both the modulator and demodulator. If the chopper is used in this way, it must be of the *short-circuiting type*; i.e., when the vibrator reed is at the center of its travel,

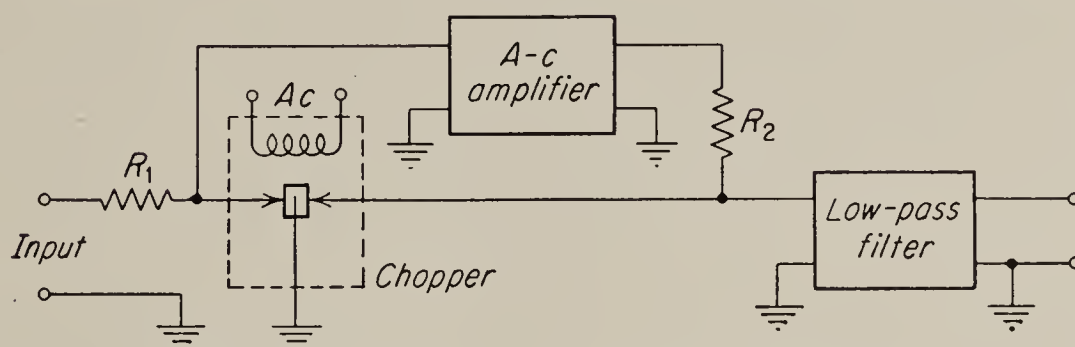


FIG. 2.35. Chopper-stabilized d-c amplifier.

all three contacts are connected together. With this type of chopper either the output or the input is always grounded. There is therefore no possibility of a feedback connection being established between output and input through the capacitance existing between the open vibrator contacts. Such feedback might result in undesirable oscillations.

The operation of the circuit is best understood by assuming the input signal to be a low-frequency sine wave and examining the wave-shapes at various points in the circuit, as shown in Fig. 2.36. Figure 2.36a shows the sinusoidal input voltage. If it is assumed that the input impedance of the a-c amplifier is infinite compared to the resistance R_1 in Fig. 2.35, the input voltage (a) appears at the input to the amplifier without attenuation when the input contact of the chopper is open; otherwise the voltage is zero. This situation is illustrated in Fig. 2.36b. Note that the amplifier input voltage still contains a component at the low signal frequency. However, if we assume that the amplifier does not pass this low frequency, then the amplifier output voltage will have the form shown in Fig. 2.36c.¹ This output will be passed to the low-pass filter only when the output chopper contact is open; hence there results the voltage shown

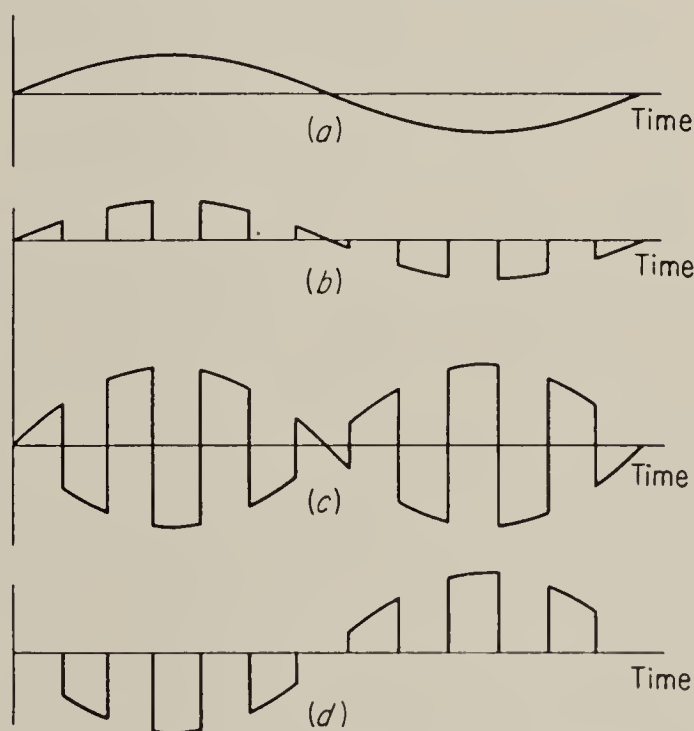


FIG. 2.36. Waveshapes of chopper-stabilized amplifier.

¹ It is shown in Chap. 6 that the amplitude of the signal component of the wave-shape of Fig. 2.36b is one-half of the input amplitude. The high-pass filters in the a-c amplifier effectively subtract this component. The reader may convince himself by performing this subtraction graphically that Fig. 2.36c is obtained.

in Fig. 2.36*d*. The low-pass filter serves to remove the chopper frequency from the output, so that the final result is ideally an amplified reproduction of the input (Fig. 2.36*a*). Note that, in the example carried through here, the amplifier was assumed to have an even number of stages, so that there is no sign reversal between the voltages of Fig. 2.36*b* and *c*. The final reversal between input and output is due to the fact that the chopper inherently applies a signal to the input at the same time that it shorts the output, and vice versa. It should be clear that an odd number of amplifier stages would have resulted in a reversal between signals *b* and *c* and no reversal between *a* and *d*.

Although the amplifier described above and others operating on the modulator principle make it possible to achieve very high gain with negligible drift, they have the very serious disadvantage that they can

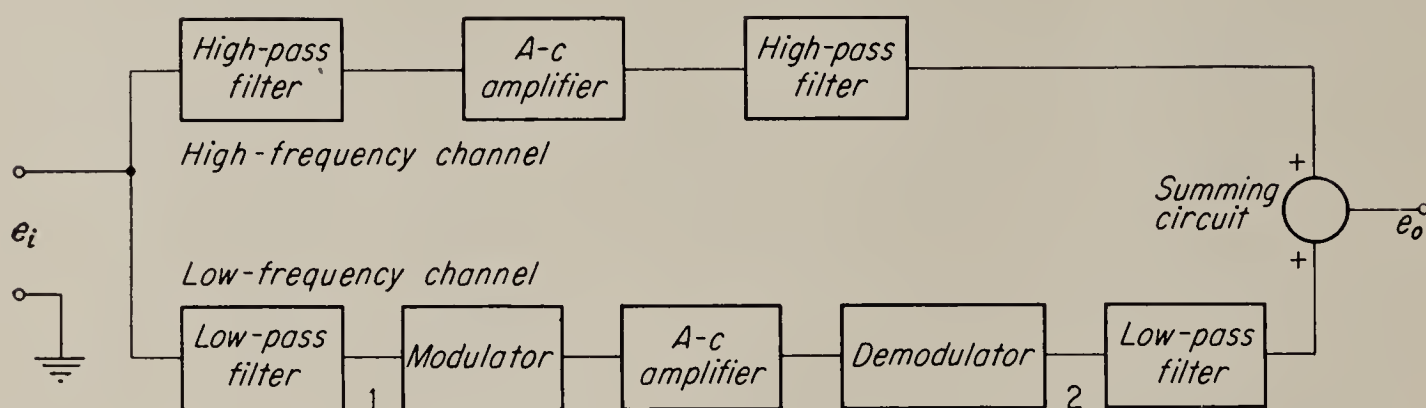


FIG. 2.37. Two-channel wideband d-c amplifier.

amplify only very low frequency signals. The reader may easily prove to himself by the method used in connection with Fig. 2.36 that, if the input frequency had been equal to the chopping frequency, the output would have been a d-c voltage. This indicates that amplification of signals at chopper frequency is impossible (see also Chap. 6). Even for signal frequencies of the order of one-tenth of the chopping frequency it would be difficult to design a low-pass filter to follow the amplifier which would remove the carrier and pass the signal without too much attenuation. The obvious means for alleviating this difficulty, i.e., using higher chopper frequencies, is limited by the fact that most choppers are built to operate at either 60 cps or 400 cps and that it is very difficult to build an electromechanical device to operate reliably at frequencies in the upper audio range. Hence, when it is desired to have high gain with zero drift over a wide frequency band extending down to zero frequency, it becomes necessary to employ a two-channel arrangement, like that shown in Fig. 2.37. Here the signal is split into high- and low-frequency components by means of the input filters shown preceding the amplifiers in the two channels. Actually, the high-pass filters shown with the high-frequency channel might be the usual RC coupling networks used in a-c amplifiers. The crossover frequency at which the signal shifts from the

high-frequency channel to the low-frequency channel must be low enough so that negligible energy at chopper frequency enters the modulator, since this will give rise to spurious signals, as indicated previously. A second low-pass filter, similar to the one shown in Fig. 2.35, is required in the low-frequency channel to remove the chopper frequency from the output. The summing circuit might take the form of the difference amplifier described in Sec. 2.11. In this case it is, however, necessary that there be a sign reversal in one channel relative to the other.

It should be noted that, unless special care is used in the design of the transfer functions of the two channels, there will be a dip in the frequency at which the signal shifts from one channel to the other. Although the transfer functions of the two channels required to avoid this will not be described here, an indication of how the problem can be solved approximately is given below in connection with Eq. 2.80. Negative feedback is

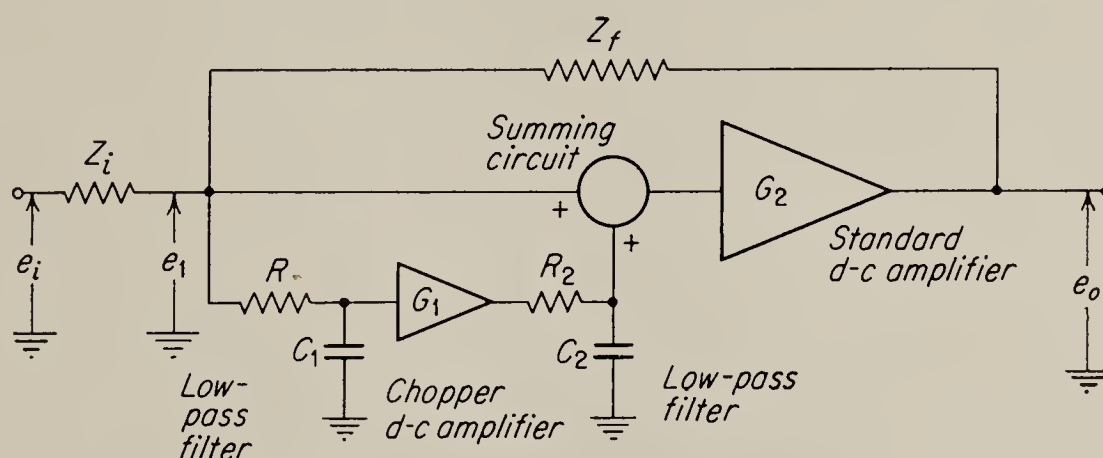


FIG. 2.38. The Goldberg d-c amplifier.

also commonly used to stabilize the gain and has the additional advantage of providing approximately constant gain for the composite amplifier even if the two channels do not have exactly the same gain. Thus it is possible to obtain an amplifier with substantially constant gain over a wide frequency range extending down to direct current, while at the same time the drift is virtually zero.

In many applications, such as the operational amplifiers used in analogue computers, it is not necessary to have constant gain at all frequencies. Large amounts of negative feedback are always employed in these amplifiers, and the primary requirement is zero drift. In this case the ideas concerning drift elimination by means of negative feedback discussed in Sec. 2.3 may be employed. The chopper-stabilized d-c amplifier of Fig. 2.37 may be used as the high-gain, driftless preamplifier. Since high gain is required only to combat drift, the high-frequency channel shown in Fig. 2.37 can be omitted and replaced by a simple bypass connection of unity gain. The summing amplifier may then be followed by several stages of d-c amplification to provide a sufficiently high loop gain at higher frequencies to ensure that the over-all gain of the amplifier

is determined by the constants of the feedback network. A simplified schematic of this arrangement is shown in Fig. 2.38, and a practical form of the circuit is shown in Fig. 2.39.¹ The bypass connection from e_1 to the summing circuit is required so that the low-pass filters in the chopper-amplifier channel do not reduce the over-all gain to very low values at

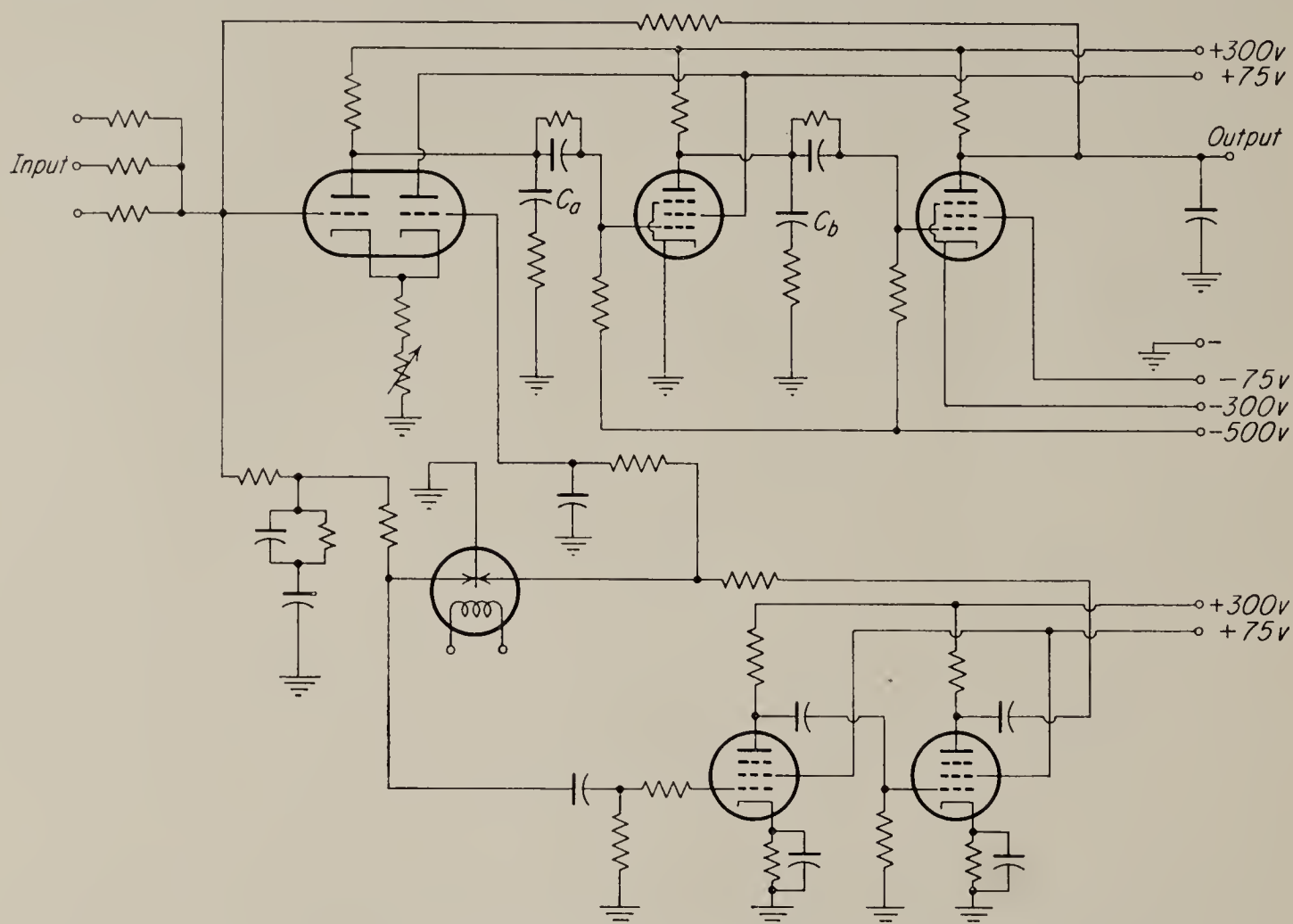


FIG. 2.39. Schematic of Goldberg amplifier.

relatively low frequencies. It ensures that the gain is at least as large as G_2 for all frequencies. The over-all forward gain is

$$\begin{aligned} \frac{\hat{e}_o}{\hat{e}_i} &= G_2 \left[1 + \frac{G_1}{(R_1 C_1 s + 1)(R_2 C_2 s + 1)} \right] \\ &= \frac{G_2 [(R_1 C_1 s + 1)(R_2 C_2 s + 1) + G_1]}{(R_1 C_1 s + 1)(R_2 C_2 s + 1)} \end{aligned} \quad (2.80)$$

When the numerator is solved for its roots, it is found that, unless the ratio of $R_1 C_1$ to $R_2 C_2$ is greater than G_1 (or less than $1/G_1$), the roots have large imaginary components, indicating a pronounced dip in the frequency response. Since this is undesirable, the ratio of the two time constants should be kept of the same order as the gain. The adjustment is not, however, very critical. The open-loop frequency response will then have the appearance shown in Fig. 2.40. Note that the gain is very high for

¹ E. A. Goldberg, Stabilization of a Wide-band D-C Amplifier for Zero and Gain, *RCA Rev.*, June, 1950, pp. 297-300.

direct current but drops to a more moderate value at intermediate frequencies. At very high frequencies the gain drops again; this is due to capacitors C_a and C_b in the d-c amplifier. These capacitors are required to make the entire feedback loop, of which the amplifier is only a part, stable. The reader is referred to standard texts on servomechanism theory¹ for details on the design of such components. The effect of these capacitors is not included in Eq. (2.80), which therefore applies only at low frequencies.

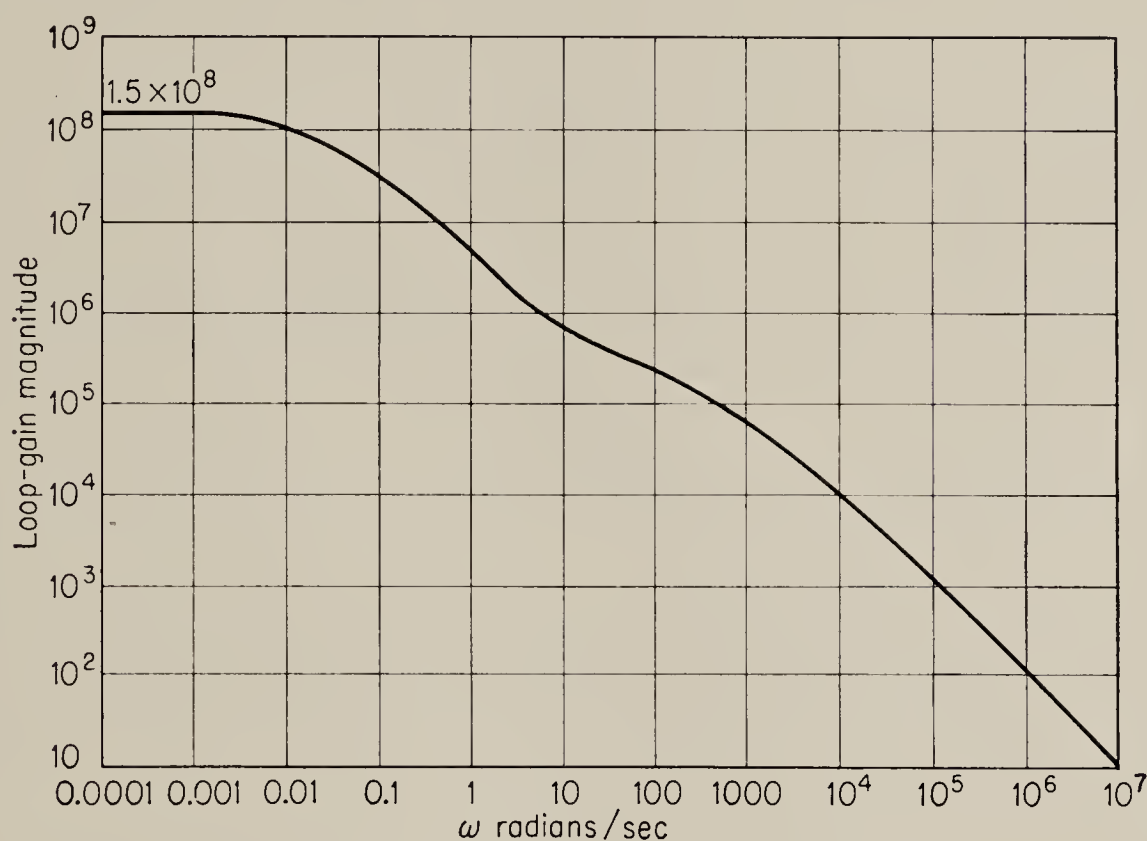


FIG. 2.40. Loop gain of Goldberg amplifier.

2.24. Design Considerations. Many design methods that are quite well established and straightforward for audio amplifiers break down when the attempt is made to apply them to d-c amplifiers. They must, therefore, often be replaced by relatively cumbersome trial-and-error methods. Thus, for instance, no really simple method is available for the optimum design of a simple triode amplifying stage having a cathode-bias resistor. Hence the following remarks have been set down chiefly to point out some of the special problems encountered in d-c-amplifier design and some of the means used to solve them. Lest the reader be discouraged by the lack of definite design procedures, it should be noted that d-c amplifiers are no more critical than audio amplifiers in their tolerance requirements, and it is usually found that large variations in component values result in only relatively minor performance changes. Hence certain standard circuits have been developed in most cases and are usually adequate.

2.25. Triode Amplifier with Resistance-coupling Network. When a simple triode amplifier is coupled to another stage, the nature of the

¹ Chestnut and Mayer, "Servomechanisms and Regulating System Design," John Wiley & Sons, Inc., New York, 1951, vol. 1.

coupling network must be considered in the design of the amplifier. Consider first the resistance coupling shown in Fig. 2.41a. The effect of this coupling on the load line for the amplifier will be examined first. In Fig. 2.41b load line 1 has been drawn for an amplifier for which $R_1 + R_2$ is infinite; the slope of the line is $-1/R_L$. The line is extended for values of negative plate voltage, even though the tube cannot operate in that region for reasons that will become apparent below.

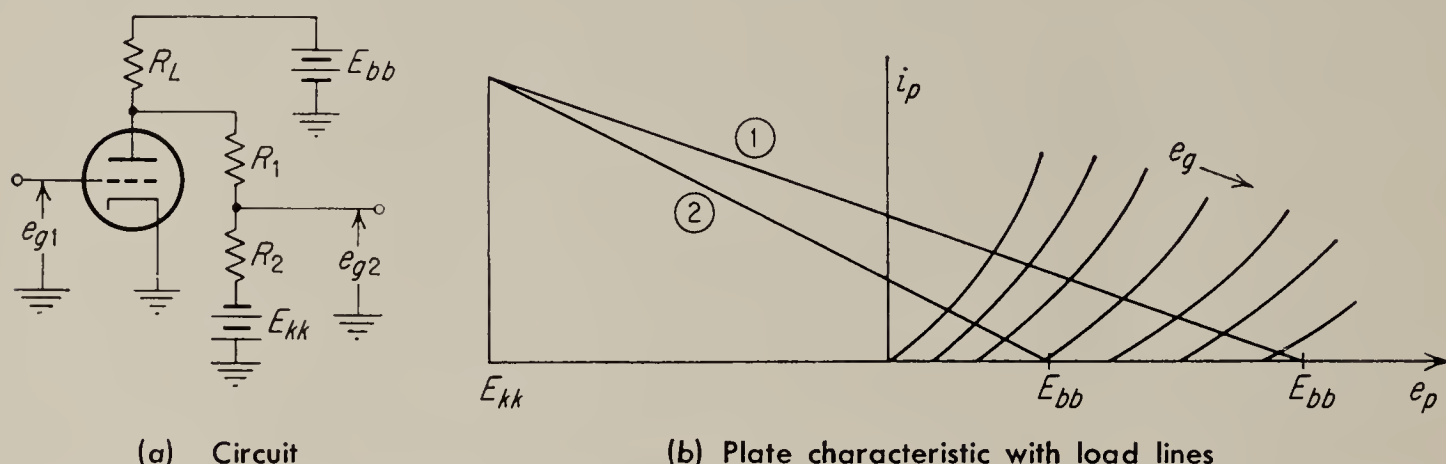


FIG. 2.41. Design of amplifier with resistance coupling: (a) schematic circuit; (b) plate characteristic with load lines.

If the resistance of the coupling network becomes finite, both the effective value of plate-load resistance seen by the tube and the effective positive supply voltage will be reduced. The plate load becomes simply the parallel combination of R_L and $R_1 + R_2$, or

$$R'_L = \frac{R_L(R_1 + R_2)}{R_L + R_1 + R_2} \quad (2.81)$$

The effective positive supply voltage is the voltage appearing at the plate of the tube when the plate current is zero. The current flowing through R_L into the coupling network causes this voltage to be less than the actual supply; thus

$$\begin{aligned} E'_{bb} &= E_{bb} - \frac{(E_{bb} - E_{kk})R_L}{R_L + R_1 + R_2} \\ &= \frac{(R_1 + R_2)E_{bb} + R_LE_{kk}}{R_L + R_1 + R_2} \end{aligned} \quad (2.82)$$

Both of these effects are indicated by line 2 in Fig. 2.41b, which represents the load line resulting from a finite $R_1 + R_2$. All the load lines must meet at the point $e_p = E_{kk}$, since there is then no voltage across R_1 and R_2 . It is clear that, if the level of impedance of the coupling network is not high enough, the tube will operate in an undesirable region with relatively small gain and be able to handle only relatively small signal swings. On the other hand, very high coupling impedance results in poor high-frequency response and increased sensitivity to noise pickup. A

compromise is, therefore, usually necessary, and the levels of R_1 and R_2 are usually taken so as to keep line 2 in Fig. 2.41b fairly close to line 1.

The design procedure in detail is as follows:

1. Pick values of R_L and $R_1 + R_2$ giving sufficient gain and signal-handling capacity.
2. Decide on operating point (usually in the middle of the linear range).
3. Specify operating voltage of next grid (usually near zero volts).
4. Find R_1 and R_2 from relation

$$\frac{R_1}{R_1 + R_2} = \frac{e_c - E_{kk}}{e_b - E_{kk}} \quad (2.83)$$

When an optimum design is required, a number of trials are usually necessary to obtain the best combination of R_L , R_1 , and R_2 .

It is theoretically possible to compensate for the poor high-frequency response of resistance-coupling networks by bridging R_1 of the coupling network with a capacitor, as shown in Fig. 2.42a. The optimum value

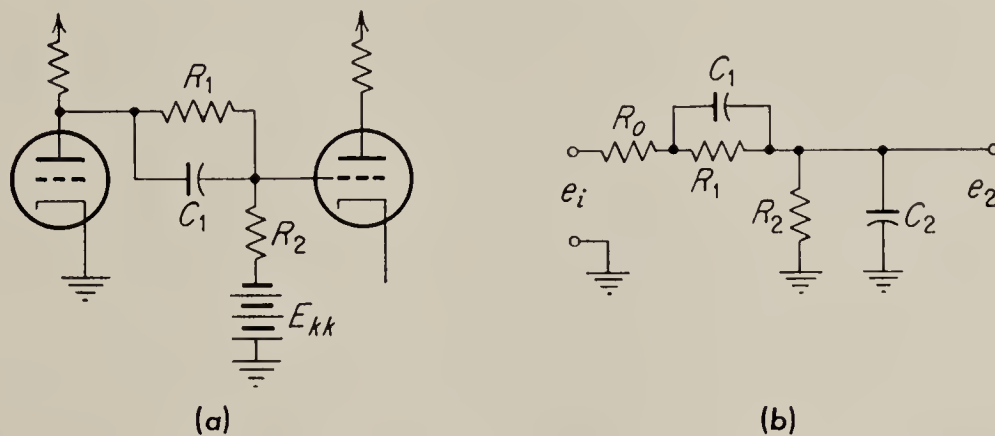


FIG. 2.42. Use of capacitor to improve frequency response of coupling network.

for this capacitor may be computed if the input capacitance of the following stage is known. This capacitance can be found from the published values of interelectrode capacitances by methods given in standard texts.¹

If this input capacitance is designated by C_2 , then the equivalent circuit of the coupling network takes the form shown in Fig. 2.42b, where R_o is the output impedance of the previous stage and C_1 the bridging capacitor. The transfer function of the coupling network is found by the methods discussed in Chap. 1. It is given by

$$\frac{\hat{e}_2}{\hat{e}_1} = \left(\frac{R_2}{R_o + R_1 + R_2} \right) \left(\frac{R_1 C_1 s + 1}{\frac{R_o R_1 C_1 R_2 C_2}{R_o + R_1 + R_2} s^2 + \frac{R_o R_1 C_1 + R_o R_2 C_2 + R_2 R_1 C_1 + R_1 R_2 C_2}{R_o + R_1 + R_2} s + 1} \right) \quad (2.84)$$

It may be shown that the best value of C_1 is the one that makes

$$R_1 C_1 = R_2 C_2$$

¹ See, for instance, Reich, "Theory and Applications of Electron Tubes," McGraw-Hill Book Company, Inc., New York, 1944, pp. 93-96.

With this value of C_1 the transfer function of the coupling network becomes

$$\frac{\hat{e}_2}{\hat{e}_1} = \left(\frac{R_2}{R_o + R_1 + R_2} \right) \left(\frac{1}{[R_o R_2 C_2 / (R_o + R_1 + R_2)]s + 1} \right) \quad (2.85)$$

i.e., the zero in the transfer function exactly cancels one of the poles, and the remaining pole is moved to a relatively high frequency. The transfer function of the coupling network without the compensating capacitor C_1 may be obtained by setting $C_1 = 0$ in Eq. (2.84). Comparison of the result of this with Eq. (2.85) indicates that the passband of the network is increased by a factor of $(R_1 + R_o)/R_o$ by use of the correct value of C_1 .

2.26. Design of Triode Amplifiers with Neon-tube Type Interstage Coupling Network. When the neon-tube coupling networks are used, the design procedure is similar to that described in the previous section.

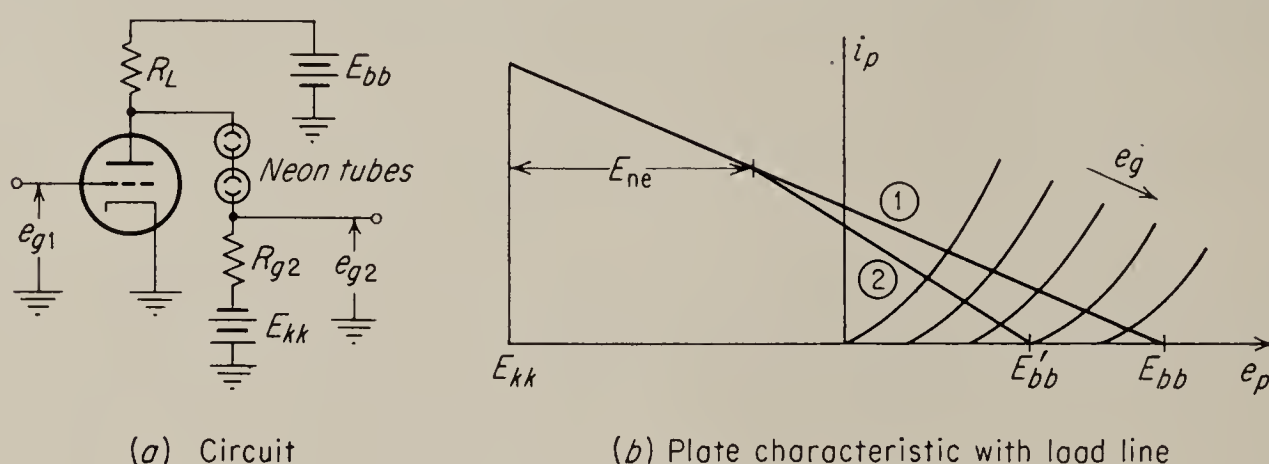


FIG. 2.43. Neon-tube coupling.

Since the operating current for neon tubes must fall within rather narrow limits, one variable, namely, the impedance level of the coupling network, is specified in advance. The effect of the coupling on the effective plate voltage and plate-load resistance is similar to that found in the resistance-coupled amplifier. The effective supply voltage is

$$E'_{bb} = E_{bb} - i_{ne} R_L \quad (2.86)$$

The effective plate-load resistance is

$$R'_L = \frac{R_L R_g}{R_L + R_g} \quad (2.87)$$

where i_{ne} is the neon-tube current and R_g the grid-leak resistance of the following stage (see Fig. 2.43a). The detailed design steps are then as follows:

1. Specify the quiescent value of the voltage at the grid of the following stage. If E_{kk} is given, this immediately specifies the value of R_g , since

$$R_g = \frac{e_{c2} - E_{kk}}{i_{ne}} \quad (2.88)$$

2. Select a value of R_L , which, together with the value of R_g decided on in (1) above, will result in an effective load line giving sufficient gain and signal range.

3. The proper number of neon tubes to use between stages is now determined by noting that the sum of the quiescent grid voltage determined in (1) plus the neon-tube drop must result in an operating point for the first stage that will permit the required signal swing in the first tube. This requirement is complicated by the fact that the voltage drop of neon tubes picked at random may vary from about 49 to 74 volts. Hence, in order to avoid selection of neon tubes, it is desirable to pick R_L in such a way that the full range of neon-tube voltages will not force the operating point of the first stage into undesirable regions.

This procedure is best illustrated by an example. Assume that e_{c2} , the quiescent second-grid voltage, is -15 volts and that a plate-signal swing of 10 volts peak is present. Let $E_{bb} = 300$ volts. The effective plate supply voltage E'_{bb} is less than this, the amount depending on the value of R'_L . Suppose two neon tubes are used. The operating-point voltage for the plate of the first stage may then vary from 83 to 133 volts with variations in neon-tube voltages (see points A and B, Fig. 2.44). The signal swing will cause the plate voltage to vary from 73 to 93 volts or from 123 to 143 volts around these two operating points. It is clear from the diagram that the lowest value of plate voltage (73 volts) would require a positive grid voltage if the lower plate-load resistance R'_{L1} were used. Hence the larger value shown in the figure, R'_{L2} , would be indicated. Another possibility is to use three neon tubes. Considerations similar to those discussed would then indicate a possible plate-voltage variation between 122 and 217 volts. This would probably be satisfactory if the plate-load resistance were R'_{L1} , but the tube would be close to cutoff with the larger plate-load resistance.

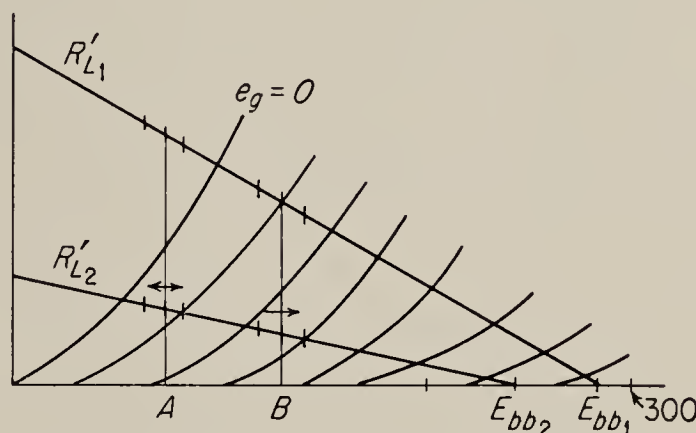


FIG. 2.44. Effect of neon-tube tolerances on operating point.

It is clear from the diagram that the lowest value of plate voltage (73 volts) would require a positive grid voltage if the lower plate-load resistance R'_{L1} were used. Hence the larger value shown in the figure, R'_{L2} , would be indicated. Another possibility is to use three neon tubes. Considerations similar to those discussed would then indicate a possible plate-voltage variation between 122 and 217 volts. This would probably be satisfactory if the plate-load resistance were R'_{L1} , but the tube would be close to cutoff with the larger plate-load resistance.

It should be noted, incidentally, that the variation of plate operating point required to accommodate random values of neon-tube voltages is accomplished by the zero balance adjustment required in every d-c amplifier. This adjustment must be designed with the proper range to handle the variations discussed here, and other possible tolerance effects. The adjustment should usually be placed in the first stage of the amplifier, so that all stages may be adjusted together.

The question of neon-tube tolerance is one of the simpler examples of

the more general problem of designing a circuit to make use of components with standard tolerances. A complete design would require a check of the effect of the tolerances of all components and a redesign of sections of the circuit which cause unsatisfactory operation under one or several extremes of component variations. Occasionally, particularly in very critical applications, it is necessary to select components specially; ordinarily, however, this should be avoided if possible.

As was pointed out in Sec. 2.17, neon-tube coupling networks have poor high-frequency response owing to the “inductive” effect of ionization time in the neon tubes. It might seem that a capacitor bridging the neon tube would improve the response here as it did with the resistance-coupling network. This procedure must, however, be applied with caution, since inspection of the circuit reveals that it has a form almost indistinguishable from the classic gas-tube relaxation oscillator. Hence, unless special precautions are taken to prevent the oscillations, capacitors across the neon tubes are not ordinarily desirable.

2.27. Design of Cathode-follower Circuits. We assume that the cathode follower operates into a specified load resistance R and that the maximum output-signal swing is to be obtained (see Fig. 2.45a for the circuit).

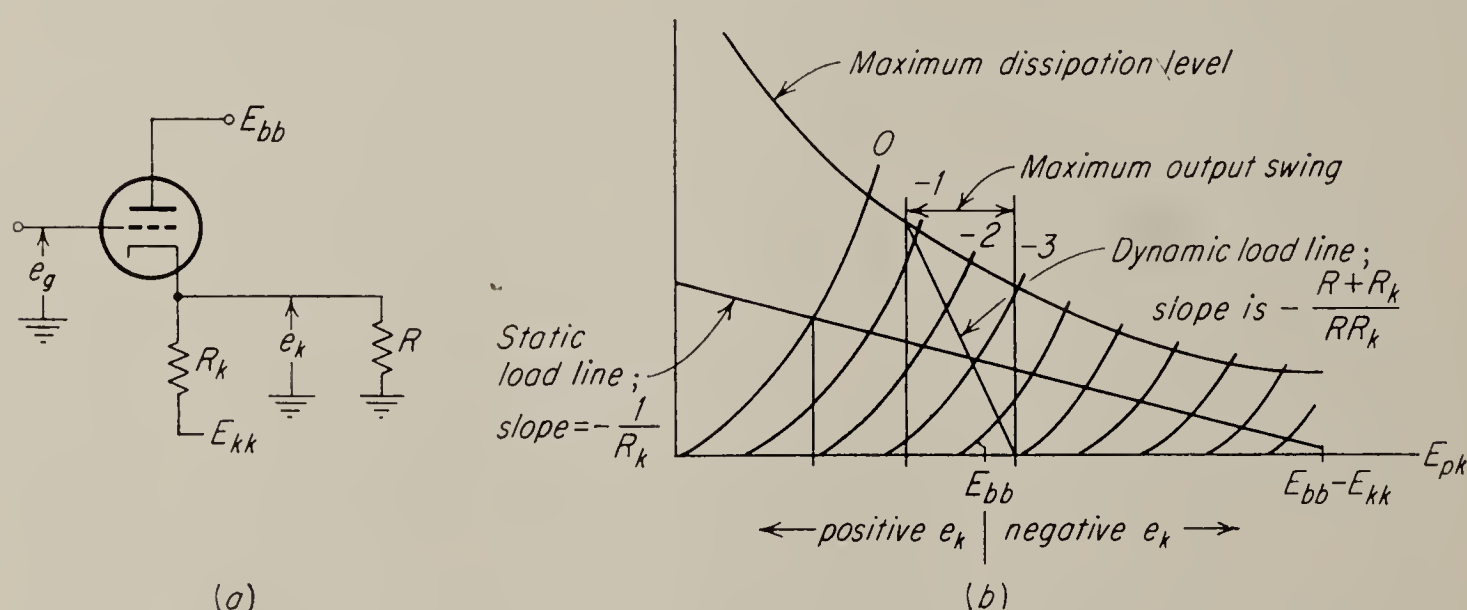


FIG. 2.45. Cathode-follower design: (a) schematic; (b) plate characteristic.

For a particular value of R_k , a “static” load line (i.e., the line for $R = \infty$) can be drawn for the cathode follower from considerations similar to those employed in amplifier stages. When the plate current is zero, the plate-to-cathode voltage is $E_{bb} - E_{kk}$. This locates one point on the static load line. When the plate-to-cathode voltage is zero, the plate current is $(E_{bb} - E_{kk})/R_k$. This locates another point and results in a line such as the one shown in Fig. 2.45b. This load line may be used to analyze the no-load behavior of the cathode follower. When the output voltage $e_k = 0$, the plate-to-cathode voltage is equal to E_{bb} . This point may therefore be considered to be the quiescent operating point.

If the load R is connected to the cathode follower, the tube operates along the dynamic load line shown in the figure. This line may be constructed as follows: when $e_k = 0$, the dynamic load line and static load line must intersect, since there is no current in the load. For other values of e_k the slope of the dynamic load line is determined by the parallel combination of R_k and R and is given by $-(R + R_k)/RR_k$.

The optimum design for the cathode resistor R_k requires a knowledge of the maximum plate dissipation of the tube. A hyperbola $e_p i_p = W_{\max}$ (the maximum plate dissipation) is drawn on the characteristic curves for the tube. The exact value of the maximum plate dissipation depends on the duty cycle to which the circuit will finally be subjected. Thus, if the output voltage is generally small and if peak voltages are required only for short times, a higher value of W_{\max} may be assumed than if the tube must deliver maximum voltages for extended periods of time. However, assuming that the maximum dissipation curve has been drawn, the optimum design is one permitting equal swings from e_k in both directions, as shown in Fig. 2.45b.

The design proceeds as follows:

1. Assuming E_{bb} , E_{kk} , and R are known, estimate an approximate value for R_k so that the static load line passes about halfway between the maximum dissipation line and the $i_p = 0$ axis. This makes possible a tentative choice for the dynamic load line, which has the slope $-(R + R_k)/RR_k$.

2. Draw the tentative dynamic load line such that equal swings on both sides of the quiescent operating point are possible (see Fig. 2.45b).

3. The static load line should bisect the dynamic load line at the operating point. If it does not, a correction must be applied to the estimated value used in step 1 and the procedure repeated until it does.

While the above procedure will result in an optimum design, it might be well to point out that optimum design is seldom justified, particularly if tube and component variations are ignored. In this case the design is much simpler, consisting in fact of step 1 above.

2.28. Design of Difference Amplifiers. Since the difference amplifier is a relatively complicated circuit, the graphical design technique using the load line cannot be applied exactly; however an approximate design that is adequate for determining limits of linear operation and quiescent operating points can be made fairly easily.

In order to establish some rough criteria for optimum values for R_L and R_k (see circuit, Fig. 2.46a), we first consider the operation of the cathode-follower section (section 2 of Fig. 2.46a), as its grid voltage is varied from a large positive value through zero to large negative values. The grid of section 1 is assumed to be grounded. If E_{c2} has a large positive value, both cathodes in following E_{c2} also take on large positive values, in fact, slightly above E_{c2} to maintain proper bias of tube 2. Tube 1 is then cut

off and presents a load of infinite impedance to the cathode follower. This mode of operation is indicated in Fig. 2.46*b* by the part of the curve marked 1. As the grid and the cathodes become less positive, the point is eventually reached where tube 1 begins to conduct. At this point it changes fairly abruptly from an infinite impedance to the relatively low value $(r_{p1} + R_L)/(\mu_1 + 1)$. The operating line of the cathode-follower section leaves the static load line and switches to the dynamic load line indicated by 2. Note that the dynamic load line is slightly curved at the top; this is due to the larger value of r_{p1} when tube 1 is just beginning to conduct. As E_{c2} is decreased further, the point is finally reached where the cathode follower cuts off, and the operating line continues along the $i_p = 0$ axis. If grid 1 had not been grounded, line 2 would have shifted either left or right, depending on whether grid 1 was positive or negative, respectively.

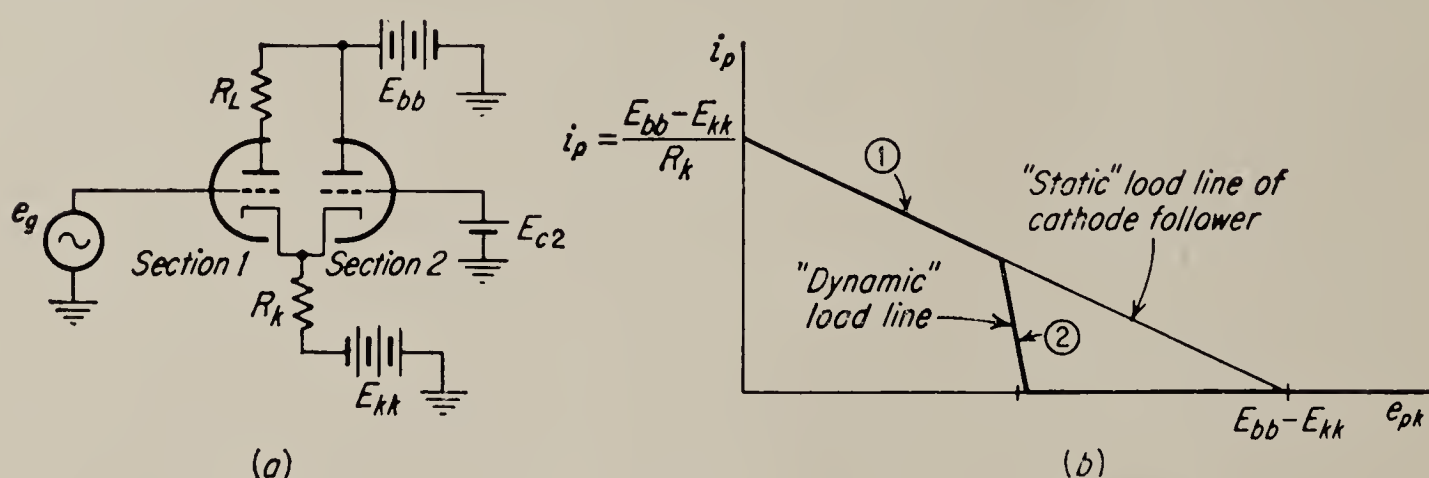


FIG. 2.46. Design of difference amplifier: (a) schematic; (b) load lines.

It should be clear that linear operation of both halves of the circuit is represented only by the line segment marked 2 in Fig. 2.46*b*. If maximum linear range is an important consideration, we should endeavor to make this line segment as long as possible. Assuming that the supply voltages are fixed, both a reduction in R_k and an increase in R_L will have this effect. Hence R_k should be just large enough to prevent excessive plate dissipation in the cathode-follower section, while R_L should be made as large as possible consistent with linear operation of the amplifier section. The operating point should then be located somewhere near the center of line segment 2 if possible, although other circuit requirements, e.g., the need to accommodate component tolerances, will usually set the operating point at some less-than-optimum value.

The design of the interstage coupling networks to follow the amplifier is not so straightforward as it was for the simple triode with grounded cathode. The reason for this is that the amplifier section of the difference amplifier has an effective cathode bias resistance equal to the output resistance of the cathode-follower section. Strictly speaking, therefore, the characteristic curves should not have the slope $1/r_p$ but rather the

slope $1/r_p' = 1/[r_p + (\mu + 1)R_o]$, R_o being the cathode-follower output resistance. Generally, however, the labor involved in replotting the characteristics from the published curves is not necessary if approximate results are sufficient, particularly since the correction is not very large. Thus an approximate design for R_L and the interstage coupling is usually carried out in the same way as that described in Secs. 2.25 and 2.26 for the simple triode. It should be noted, however, that the correct value of gain cannot be deduced directly from the construction of a load line on the published tube characteristics. The procedure should rather be to determine values for μ and r_p at the operating point and to compute the gain from Eq. (2.37).

One final problem that should be considered in connection with this circuit is the determination of the proper value of voltage on grid 2, once an operating point for plate 1 has been specified. The reader should be cautioned here, incidentally, against the fallacy of assuming that the two grid voltages will have approximately equal quiescent values simply because the circuit amplifies the voltage difference existing between the grids. Such is not the case, primarily because the circuit is not symmetrical.

When the quiescent value of e_{p1} is given, the procedure for finding the corresponding value of E_{c2} is as follows: (We assume E_{c1} to be zero for the sake of simplicity, but the method is easily extended to other values of E_{c1} .)

1. With e_{p1} and R_L known, the plate current i_{p1} and the required grid bias of section 1 may be read from the published plate characteristics of the tube. Hence, if the grid is grounded, the cathode must be positive by the amount of the bias.

2. Once the cathode voltage has been determined, the voltage across the cathode resistor is known. Hence the current in the cathode resistor can be computed.

3. The plate current in the cathode follower is equal to the difference between the current in the cathode resistor and the plate current of section 1.

4. Since the plate of the cathode follower is at E_{bb} and since the cathode voltage is known from step 1, the plate-to-cathode voltage is known. The plate current is known from step 3. These two ordinates establish the required grid bias in the cathode follower.

5. The desired voltage on the cathode-follower grid is then the sum of the cathode voltage (from step 1) and the (negative) grid bias from step 4. Usually this voltage is negative with respect to the voltage on grid 1.

In addition to clarifying circuit operation these steps make possible the determination of the size of the potentiometer required to adjust the circuit to accommodate various component tolerances.

PROBLEMS

2.1. A cathode follower using a 12AX7 vacuum tube has a cathode resistor of 200 kilohms; $E_{bb} = 300$ volts; $E_{kk} = -300$ volts. Vibrations of the grid cause μ to change by ± 5 per cent; the change of plate resistance is negligible. With grid grounded, find the change of output voltage caused by the vibrations.

2.2. Compute the drift due to variations of E_{bb} and E_{kk} for the difference amplifier (Fig. 2.11). After obtaining the exact expression, simplify the result by assuming that the μ 's and r_p 's of the two tube sections are identical and that $R_k \gg r_p$.

2.3. Show that the drift in a properly adjusted Miller circuit (Fig. 2.13) due to variations of E_{bb} is given by Eq. (2.50) if the assumption is made that the μ 's of the two tube sections are identical.

2.4. Figure 2.47 shows a circuit that may be used to add three signals. Find the expression for e_o as a function of e_1, e_2, e_3 . Assume that all three tubes have the same μ and r_p and that $R_k \gg r_p$.

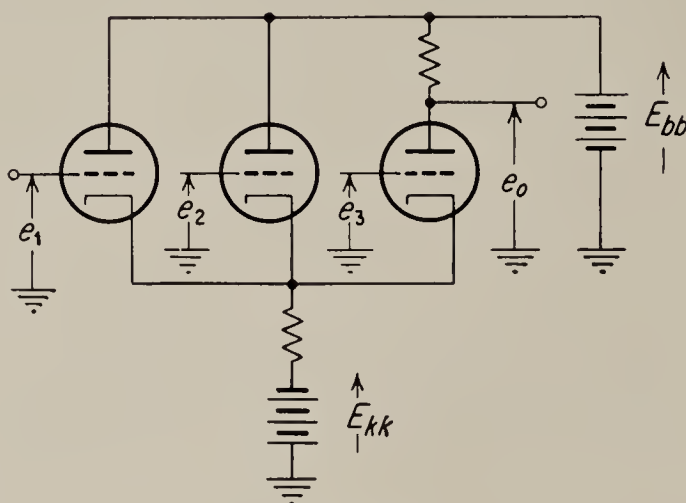


FIG. 2.47

2.5. (a) Determine the drift caused by a variation of E_{bb} on the output of the circuit in Prob. 2.4. (b) Repeat for a variation of E_{kk} . (c) Repeat for a change in heater voltage.

2.6. (a) Find the gain and output impedance of the circuit of Fig. 2.48. Assume the two tubes are identical. (b) Find the drift due to change of E_{bb} . (c) Find the drift due to change in heater voltage.

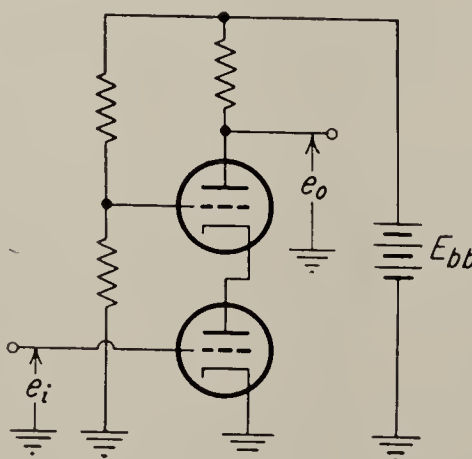


FIG. 2.48

2.7. The circuit (Fig. 2.49) shows a series-type amplifier which is supposed to have the feature that drift due to heater-voltage variation is canceled in it if R is correctly adjusted. (a) Find the value of R that provides perfect cancellation of filament-

voltage variation effect. (b) Determine the drift due to changes in E_{bb} . (c) Determine the gain e_o/e_i . Assume that the tubes are identical.

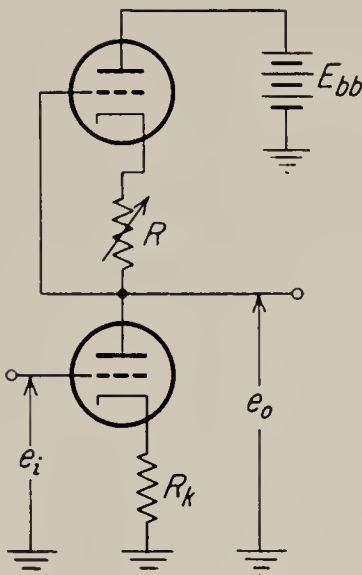


FIG. 2.49

2.8. When transistor circuits are to be analyzed by means of the node method the transistor equations are most conveniently expressed in terms of y parameters as follows:

$$\begin{aligned} i_1 &= y_{11}v_1 + y_{12}v_2 \\ i_2 &= y_{21}v_1 + y_{22}v_2 \end{aligned}$$

Express the four y parameters in terms of the h parameters.

2.9. Show that the equivalent circuit shown in Fig. 2.50 approximately represents the effect of I_{co} in the grounded-emitter circuit.

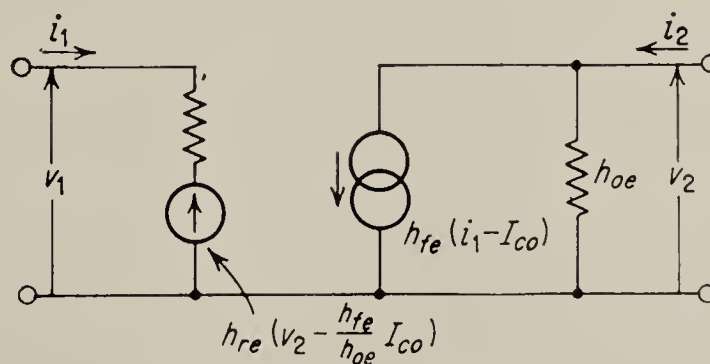


FIG. 2.50. Equivalent circuit of common-emitter transistor amplifier showing effect of I_{co} .

2.10. Obtain the h parameters for the grounded-base transistor amplifier in terms of the common-emitter parameters [see Eq. (2.79)]. Use the approximations that $h_{re} \ll 1$ and that $h_{ie}h_{oe} \ll h_{fe}$.

2.11. Show that the approximate expressions for the voltage gain, input impedance, and output impedance of the grounded-collector transistor amplifier as given in Table 2.1 are correct.

2.12. Compute the modified h parameters for the common-emitter transistor amplifier with a resistance R_e connected between emitter and ground. Find the effect of R_e on the input impedance voltage gain and current gain. Assume that $R_e h_{oe} \ll 1$.

2.13. Repeat Prob. 2.12, but consider a resistance R_f connected between collector and base. Assume that $R_f \gg h_{ie}$.

2.14. Find the gain $v_2/(v_{i1} - v_{i2})$ for the difference amplifier of Fig. 2.31. Also find the input impedance seen by v_{i1} .

2.15. Find the drift in v_2 due to change in I_{co} for the difference amplifier of Fig. 2.31.

2.16. The circuit of Fig. 2.51 shows a resistance-coupled d-c amplifier with the following characteristics:

$$\begin{aligned}\text{Tube type} &= 6\text{SL7} \\ R_L &= 300 \text{ kilohms} \\ E_{bb} &= 300 \text{ volts} \\ E_{kk} &= -300 \text{ volts}\end{aligned}$$

It is desired to optimize the gain and the gain-bandwidth product as a function of the resistance level of the coupling network. For this purpose the gain and gain-bandwidth product are to be computed for $R_1 + R_2 = 2$ megohms, $R_1 + R_2 = 1$ megohm,

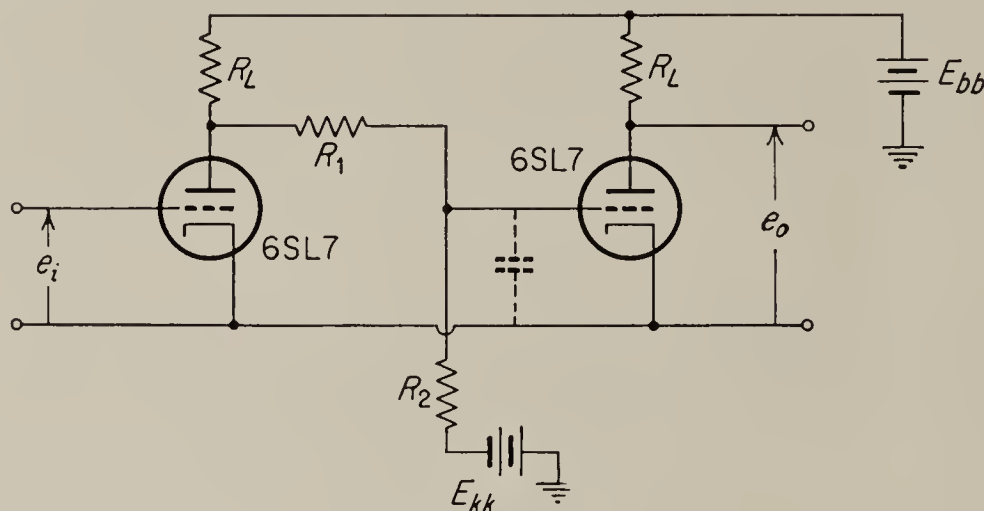


FIG. 2.51

and $R_1 + R_2 = 0.5$ megohm. For the purpose of this problem the bandwidth is defined as the frequency in radians per second for which the gain of the amplifier is 0.707 times the value at direct current. The input impedance of the second stage of the amplifier may be assumed purely capacitive with a value of $150 \mu\text{mf}$. Assume further that the quiescent value of the grid voltage for both tubes is -1 volt, that the output impedance of e_i is zero, and that the load impedance seen by the second stage is infinite.

Sketch a curve of the behavior of the gain and the gain-bandwidth product as a function of $R_1 + R_2$. Comment on the result.

2.17. The d-c amplifier shown in Fig. 2.52 is to be designed to permit a maximum load-voltage swing of ± 50 volts into the 10-kilohm load. Quiescent value of input and output is 0 volt. Find values for R_1 , R_2 , R_3 , and R_4 ; determine the number of neon tubes required between the stages and the required adjustment range for e_{g2} .

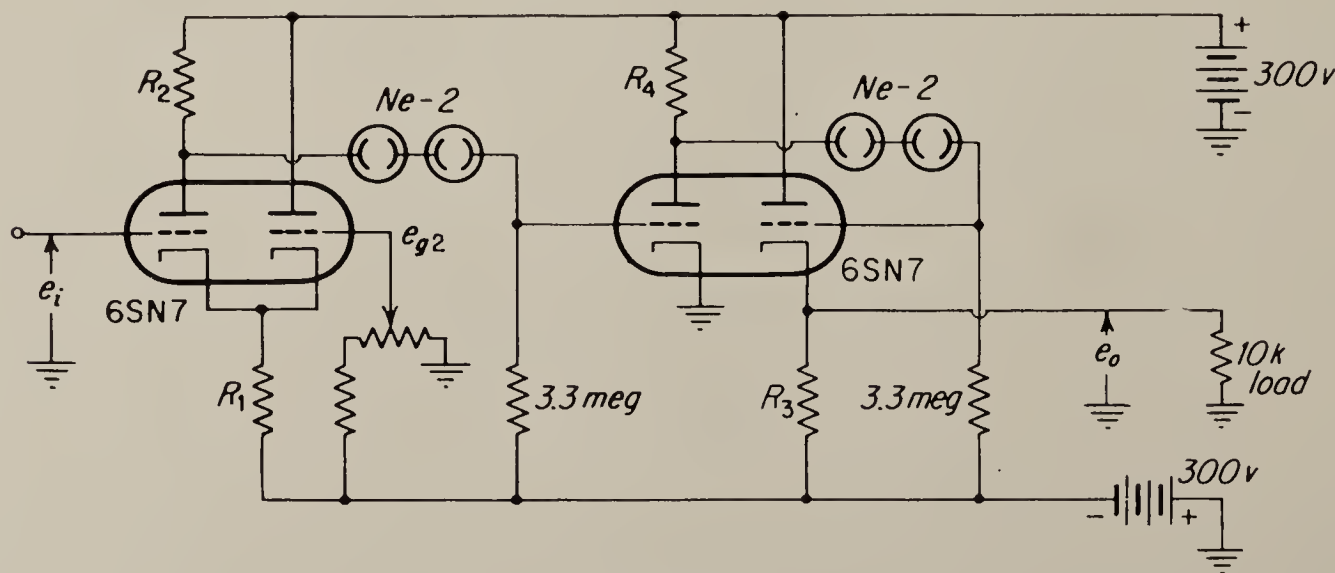


FIG. 2.52

CHAPTER 3

POWER AMPLIFIERS

3.1. Introduction. Usually the power element or output stage of a closed-loop system requires a driving source capable of supplying considerable power itself. In this chapter two convenient types of power amplifiers, the thyatron amplifier and the relay amplifier, are considered in some detail. Also, the increasingly popular transistor power amplifier and the magnetic amplifier are discussed.

Rotating amplifiers will be considered in Chap. 4 of this text. The design of conventional a-c amplifiers has been the subject of exhaustive discussion in the literature for many years and will not be covered here in any detail. A paragraph on the subject has been included in this chapter essentially as an introduction to the literature. D-c amplifiers have been discussed in Chap. 2 of this text, and the paragraph in this chapter extends this material to d-c power amplifiers.

3.2. A-C Power Amplifiers. Most textbooks on electronics consider the conventional power amplifier in considerable detail.¹ Class A, class AB, and class B single-ended and push-pull vacuum-tube-amplifier design methods are straightforward and may be as analytical as the designer desires. Class C amplifiers are usually designed by graphical methods² because of the difficulty of analytical techniques. Plate efficiency and power output for a given tube increase, in order, as it is operated as class A, AB, B, or C. Single-tube operation is practical with class A, but push-pull operation is desirable if the power level is appreciable. With the other classes, push-pull operation is required, unless the load is a tuned circuit. An example of a single-ended class A application in a control system might be the driving amplifier for an instrument synchro or resolver.³

It will be remembered by the reader that the various classes of oper-

¹ See, for example, Reich, "Theory and Applications of Electron Tubes," McGraw-Hill Book Company, Inc., New York, 1944, chaps. 7, 8.

² See, for example, Members of the staff of the Department of Electrical Engineering, Massachusetts Institute of Technology, "Applied Electronics," John Wiley & Sons, Inc., New York, 1943, chap. 10, art. 4.

³ Valley and Wallman, "Vacuum Tube Amplifiers," Radiation Laboratory Series, vol. 18, McGraw-Hill Book Company, Inc., New York, 1947, chap. 9.

ation of vacuum-tube amplifiers are defined by the grid bias. For class A operation the grid is biased so that plate current flows throughout the input-signal voltage cycle. In class B operation the grid is biased to cutoff so that plate current flows only when the grid is driven in the positive direction by the grid signal. Thus if the input signal is a sine wave, the plate current flows for one half the cycle. Class AB operation is intermediate between these first two. In class C operation the grid is biased beyond cutoff, usually to at least twice the cutoff value. As is readily seen, the transfer from signal voltage at the grid to plate current is grossly nonlinear except in class A operation. If desired, the subscript 1 or 2 may be added to the designation of the class of operation to denote the magnitude of grid signal anticipated. If the grid is driven positive with respect to the cathode, thus drawing grid current, the subscript 2 is added; otherwise the subscript 1 is used.

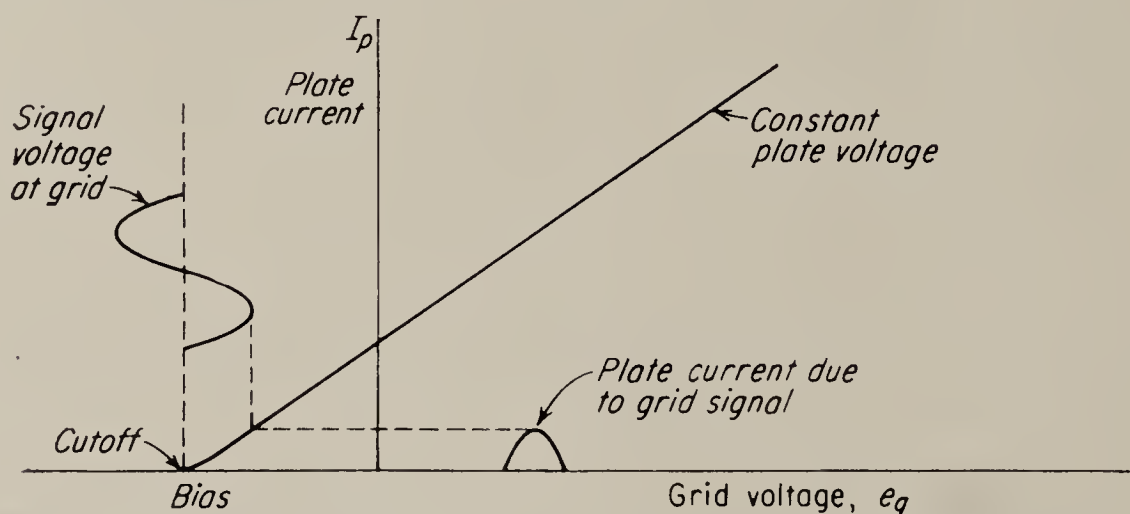


FIG. 3.1. Class B operation as shown on a typical transfer characteristic, which shows plate current plotted against grid voltage, with the voltage from plate to cathode of the tube held constant. The grid-signal amplitude shown is causing class B₁ operation.

In *push-pull* operation, two tubes are required. The two tubes are driven by signal voltages that are 180° out of phase. Thus when one tube is nearing cutoff, the other tube is entering its operating range. The resultant composite transfer is linearized, and distortion is reduced. Although the theoretical values are somewhat higher, under actual operating conditions the plate efficiency of class A amplifiers is about 30 per cent; of class B amplifiers, about 60 per cent; and of class C amplifiers, 70 to 80 per cent. Thus it may be seen that it is usually worth while to employ class B operation, even though the design procedure is somewhat more involved than that for class A amplifiers.

The definition of class B operation requires that the tube be biased to cutoff, as shown on the transfer characteristic in Fig. 3.1. Since plate current flows for only one half of the grid-voltage cycle, the resultant plate-current wave is seriously distorted. A second tube is added to the circuit that will operate on the other half of the signal-voltage wave, and the resultant plate currents are added in the load. In a-c applications

the load is usually coupled to the amplifier by means of a transformer. The transformer changes the impedance of the load as seen by the tubes in such a way as to allow a maximum transfer of power. It is conceivable that, in servomotor applications, a single tube could be used in class B operation, since the motor will respond primarily to the basic carrier frequency. However, this type of operation would cause saturation of the coupling transformer by the d-c component of current flowing in its primary. In addition, harmonics of the carrier would flow in the motor, resulting in power losses and possible spurious torques and saturation effects. Figure 3.2 shows a simplified push-pull power-amplifier circuit.

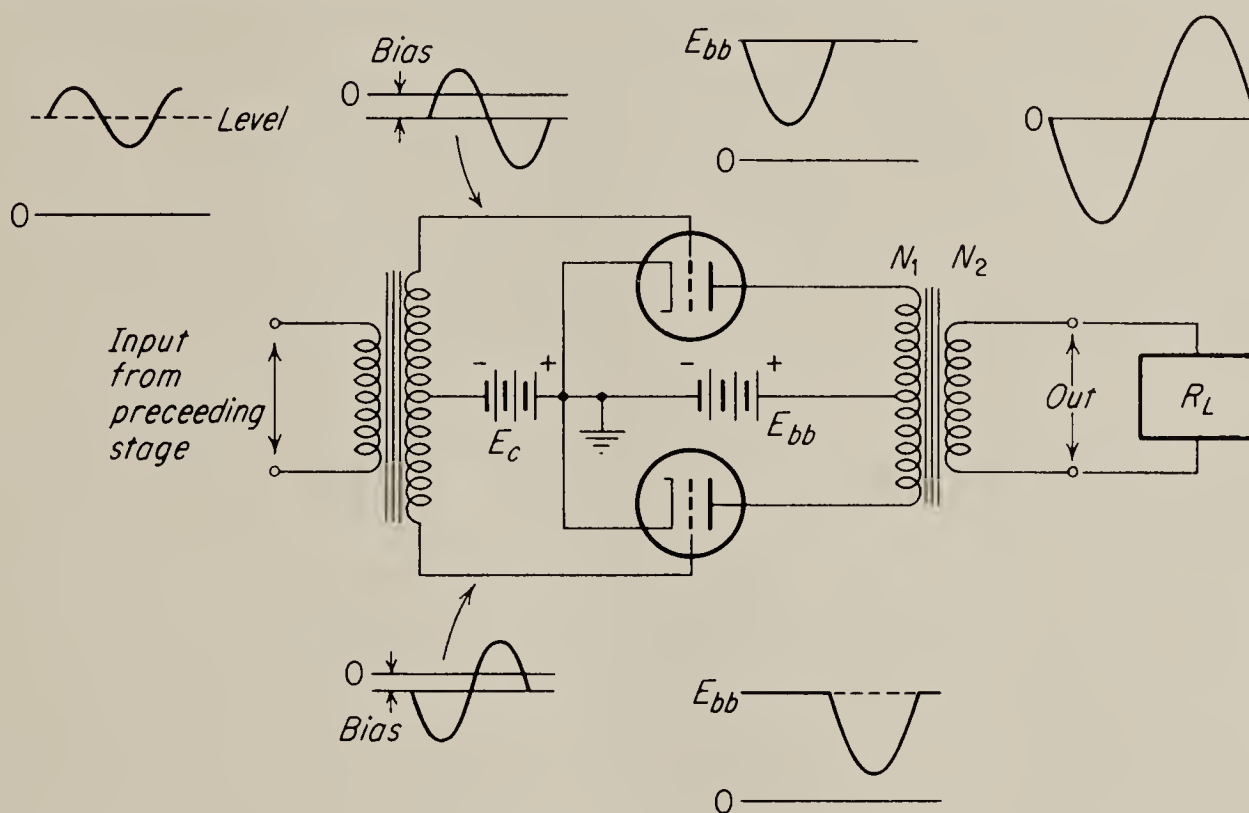


FIG. 3.2. Push-pull amplifier circuit.

A signal on the input transformer drives one grid positive and the other grid negative. The plate currents flow in opposite directions in the output transformer, thus canceling any d-c component of flux in the transformer core. In Fig. 3.3 the composite plate characteristics of the tubes are shown. Composite plate characteristics are commercially available for tubes such as the 6L6 and 807 that are commonly used in push-pull circuits. The proper plate-battery voltage and grid-bias voltage must be carefully chosen to establish the proper operating point. In Fig. 3.3 the curvature of the characteristics at low plate current has been exaggerated. We note that, strictly speaking, this is not class B operation, since both tubes conduct for small signal voltages. If the characteristics are adjusted so that the tubes are completely cut off, the composite characteristics are no longer straight lines and a certain amount of distortion is introduced.¹ If we assume that the characteristics are essentially straight lines, it is possible to derive an equivalent circuit for the oper-

¹ Reich, *op. cit.*, art. 8.1.

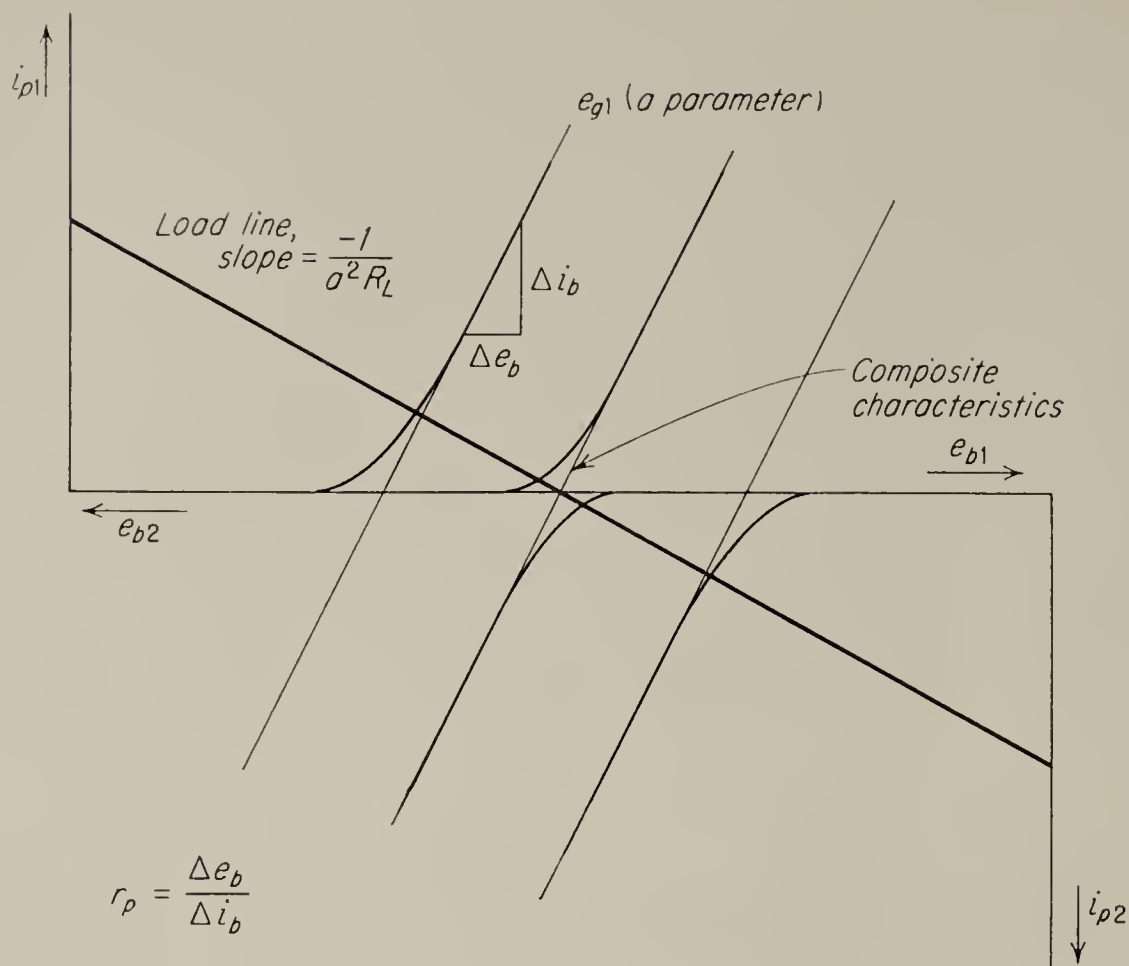


FIG. 3.3. Composite plate characteristics for push-pull operation. Strictly speaking, this is class AB operation.

ation in the class B mode. Essentially one tube is open-circuited, while the other tube is conducting. We may represent this as an a-c equivalent circuit, as shown in Fig. 3.4. As usual, μ is defined as the ratio of a change

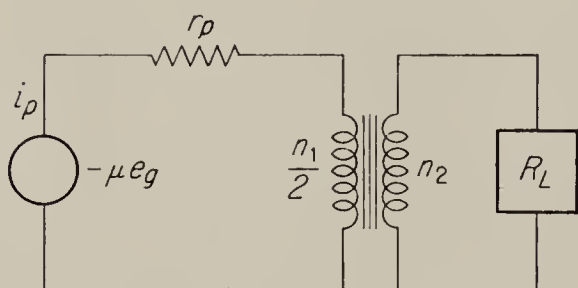


FIG. 3.4. A-c equivalent circuit for class B push-pull amplifier.

of grid voltage to the change of plate voltage required to maintain constant plate current, and r_p is the ratio of Δe_p to Δi_p at a constant grid voltage. Note that only one half of the turns of the primary of the output transformer are utilized by the plate current of either tube. Thus the equivalent turns ratio of the transformer must account for this

fact. The load impedance may be transferred to the primary by the turns ratio squared; thus

$$\frac{i_p}{e_g} = \frac{-\mu}{r_p + a^2 R_L} \quad (3.1)$$

where a is defined as

$$a \triangleq \frac{n_1}{2n_2} \quad (3.2)$$

The power dissipated in the load is

$$P_{\text{load}} = I^2 R = \left(\frac{-\mu e_g}{r_p + a^2 R_L} \right)^2 a^2 R_L \quad (3.3)$$

For e_g , μ , and R_L constant we may determine the optimum turns ratio by differentiating load power with respect to a and setting the result to zero. The result is

$$a^2 = \frac{r_p}{R_L} \quad (3.4)$$

which is the familiar statement that the load impedance must be matched to the impedance of the generator for maximum power transfer. Occasionally one is led by this familiar result into assuming that the converse is true. The reader may show that, if the internal impedance r_p of the generator can be adjusted by choosing the proper tube, it should not be made equal to the load impedance but rather should be made as small as possible, for maximum power transfer.

In general, the load is not purely resistive. If the load consists of a servomotor, for instance, it will have a large inductive reactance. Under these circumstances the equivalent circuit will be as shown in Fig. 3.5. The load impedance is referred to the primary of the output transformer. For this type of load the plate current is

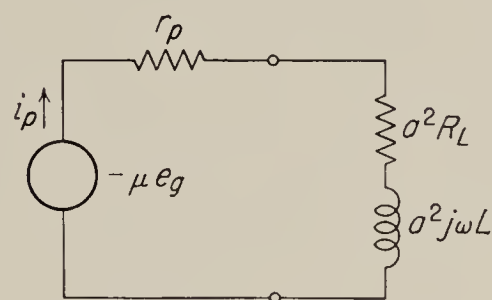


FIG. 3.5. Equivalent circuit with reactive load.

$$-\mu e_g = r_p i_p + a^2 R_L i_p + a^2 L \frac{di_p}{dt} \quad (3.5)$$

The plate current is no longer in time phase with the generator voltage,

and no longer does the simple load line describe the operation of the circuit. From Eq. (3.5) we see that only when the rate of change of plate current is zero will the operating point lie on the load line. When the plate current is increasing, the voltage across the inductance opposes the increase, and when the current is decreasing, the inductance tends to maintain the flow. It can be shown¹ that, for a sinusoidal input signal, the operating point follows an ellipse whose major axis is inclined some-

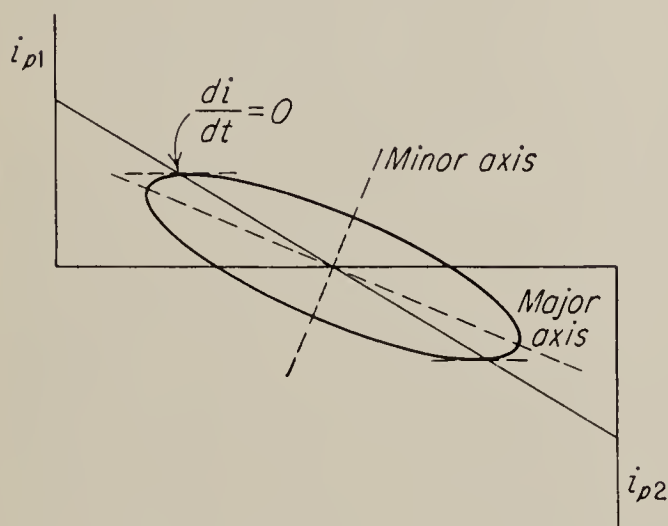


FIG. 3.6. Operating locus for inductive load.

what to the load line, as shown in Fig. 3.6. The ratio of the distance along the major axis to the minor axis is equal to the ratio of the load

¹ Dow, "Fundamentals of Engineering Electronics," John Wiley & Sons, Inc., New York, 1945, p. 274.

resistance to the load reactance. The presence of inductive reactance in the load reduces the maximum power that can be delivered to the load, while at the same time the required volt-ampere capacity of the driving amplifier remains high.

In a-c servo applications the frequency of the input signal is normally a narrow band about the carrier frequency of the system. It is therefore possible to correct the power factor of the load by placing a capacitor across it. The tuned load will absorb more power, thus more effectively utilizing the capacity of the driving amplifier. At the same time, since the loads encountered in servo applications normally have a large resistive component, there is little possibility of making the tuning so sharp as to reject the sidebands of the carrier. The resistive component also limits the usefulness of class C amplifiers for this application. Usually the effect on the phase lag of the control loop by filters at the carrier frequency must be considered.¹

Power transistors of the junction type have been under extensive development, and units capable of handling currents of over 10 amp at collector voltages of about 100 volts are presently available commercially. In addition to the well-known general advantages of transistors over tubes, transistor power amplifiers have several special advantages. First, practical transistor amplifiers can achieve almost the theoretical maximum power efficiency for a given class operation. Efficiencies of 48 per cent are achieved in class A operation and 70 per cent in class B operation.² Second, the power output of a transistor is typically in the form of low voltage and high current. Since many control-system devices such as solenoids, relays, motors, magnetic amplifiers, etc., require current for torque or flux, this leads to an excellent match between the transistor and its load in many control-system applications.

The power-handling capacity of a transistor is limited by the temperature at which the collector junction can operate and by the breakdown voltage of the collector. In certain connections a condition of temperature runaway can occur. The quiescent collector current raises the temperature of the collector junction; this causes an increase in the collector current. Under normal loads and ambient temperatures the effect is not cumulative, but under heavy currents and increased ambient temperature, compensating circuits must be considered.

As with vacuum tubes, the class B push-pull transistor power amplifier is most common. Its high efficiency and low distortion are its main advantages, since even harmonics are canceled. Of the three possible

¹ See Truxal, "Automatic Feedback Control System Synthesis," McGraw-Hill Book Company, Inc., New York, 1955, sec. 6.7.

² R. A. Hilbourne and D. D. Jones, Transistor Power Amplifiers, *Proc. Inst. Elec. Eng.*, vol. 102, part B, pp. 763-774, 1955.

symmetrical connections, the common-base connection is least used. Its power gain is low and decreases for large signals. The circuit's main advantage is its low distortion. The power gain of the common-collector circuit is typically several times higher than the grounded-base connection, and its distortion is low. This connection has some application in high-quality audio amplifier circuits and in circuits in which maximum power output is more important than power gain.

The common-emitter connection shown in Fig. 3.7 will typically have 16 to 40 times more power gain than the other circuits, and its somewhat

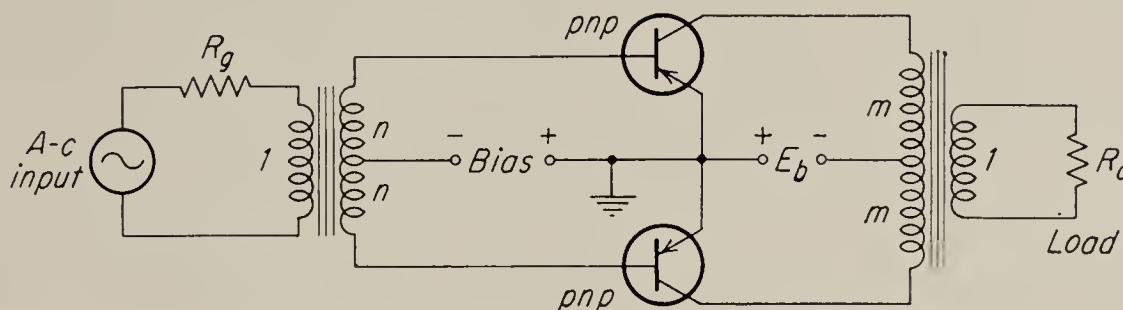


FIG. 3.7. Basic circuit of the grounded-emitter push-pull amplifier with p - n - p junction transistors.

greater inherent distortion is not usually an important drawback in control systems. In many control-system applications the load is center-tapped, and the output transformer may be eliminated.

The power gain of the common-emitter circuit is¹

$$G_p = \left(\frac{\alpha_0}{1 - \alpha_0} \right)^2 \left(\frac{1}{1 + 3B} \right)^2 \frac{4R_L}{r_s} \quad (3.6)$$

where α_0 = current gain for zero emitter-current bias

B = fractional third-harmonic distortion

R_L = reflected load resistance ($m^2 R_o$)

r_s = reflected input resistance ($n^2 R_g$)

where m , n , R_g , and R_o are defined in Fig. 3.7. It will be remembered that the theoretical maximum for the current gain α of a junction transistor is unity. Typical operating values are between 0.94 and 0.98. B , the third-harmonic distortion, is usually between 0.1 and 0.2.

The design of transistor circuits cannot rely heavily on the load-line techniques used in vacuum-tube amplifiers, since the collector characteristic, like that shown in Fig. 3.8, depends on the source impedance. From the general four-terminal equations given by Shea² and others,

$$\begin{aligned} v_1 &= r_{11}i_1 + r_{12}i_2 \\ v_2 &= r_{21}i_1 + r_{22}i_2 \end{aligned} \quad (3.7)$$

¹ *Ibid.*

² Shea, "Principles of Transistor Circuits," John Wiley & Sons, Inc., New York, 1953, p. 34.

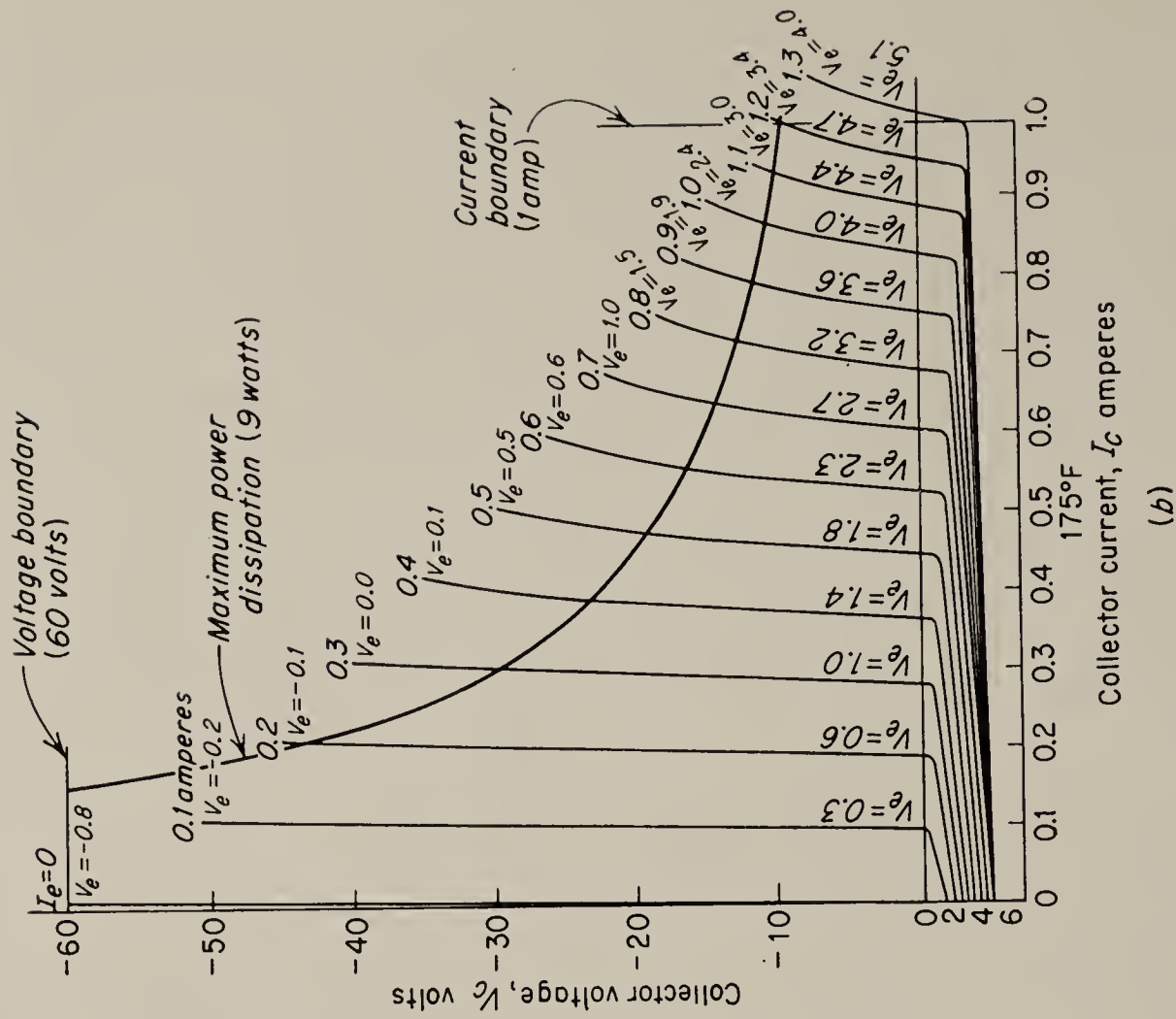
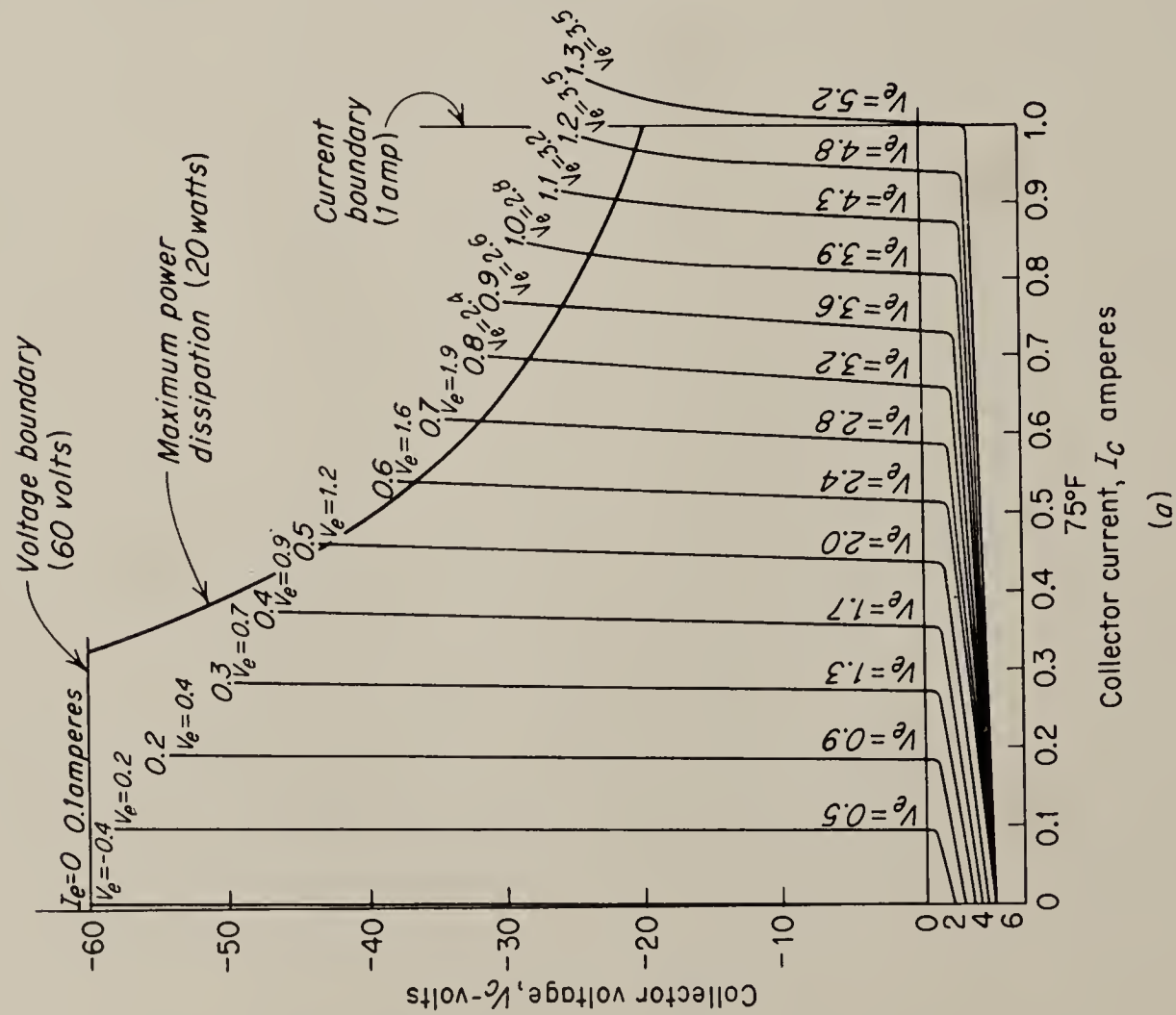


FIG. 3.8. Typical static collector characteristic of a junction transistor.

we may find the output impedance

$$R_o = r_{22} - \frac{r_{12}r_{21}}{r_{11} + r_s} \quad (3.8)$$

Thus r_s has a very definite effect on the characteristic. Fortunately the typical junction-transistor characteristic is so nearly linear that good results are obtained by determining the small-signal parameters and employing, even for large signals, the small-signal equivalent circuit shown in Fig. 3.9. The characteristic curves are used principally to check operating points and voltage and power limits. In Fig. 3.10 are shown the composite collector characteristics of the 2N57 *p-n-p* junction transistor with the operating limits set by the manufacturer.

3.3. D-C Power Amplifiers. The problems involved in d-c power amplifiers are the same problems discussed in Chap. 2. However, since the voltage gain of a power amplifier is usually relatively low, noise and

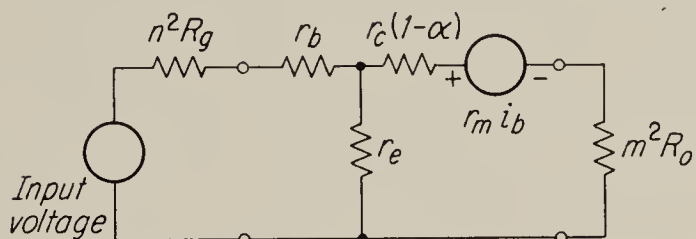


FIG. 3.9. Equivalent circuit of class B grounded-emitter push-pull amplifier.

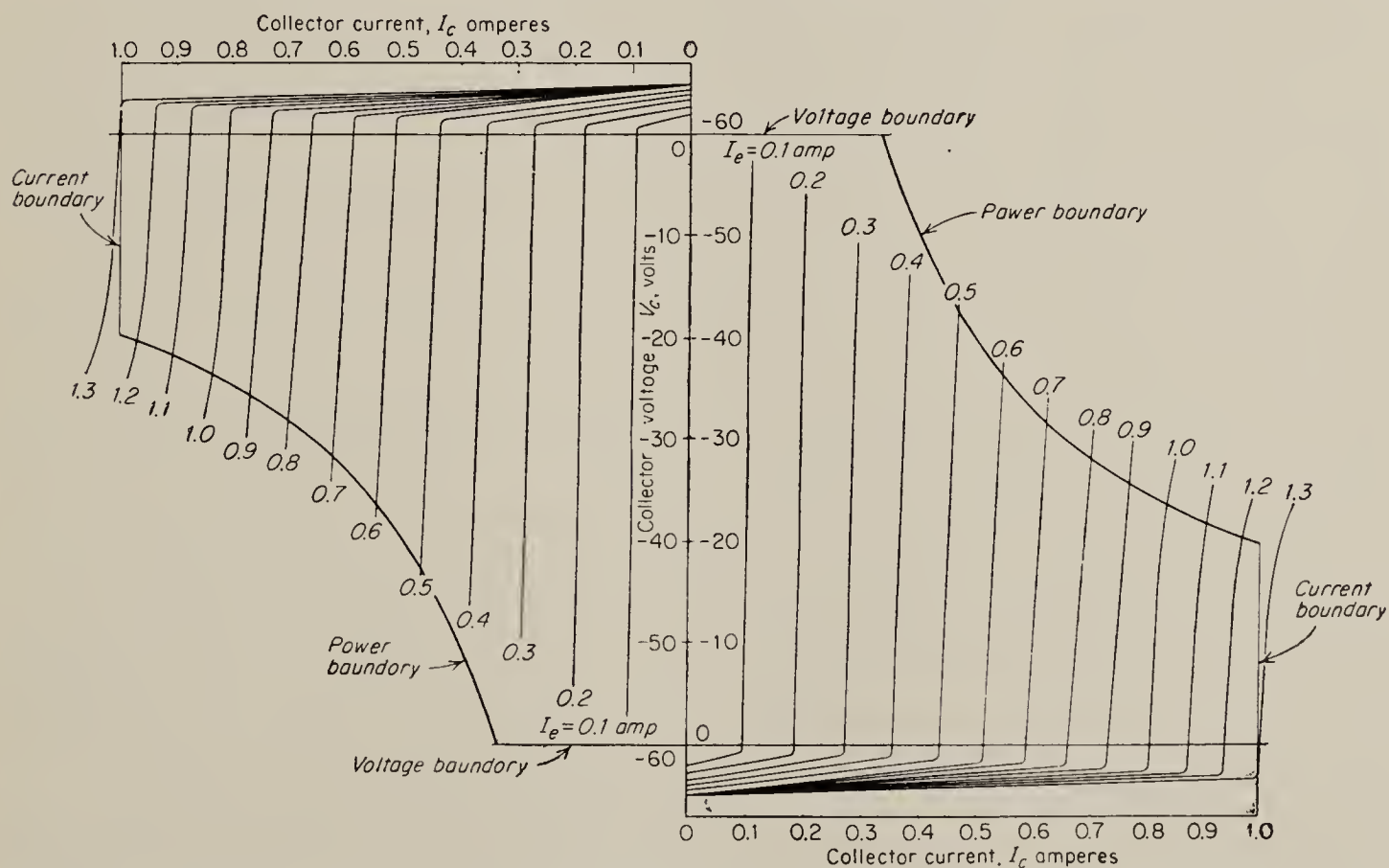


FIG. 3.10. Push-pull composite characteristics for class B operation. The two curves are matched at the operating point $V_c = -60$ volts.

drift introduced in this stage do not have so serious an effect as noise and drift introduced at lower signal levels. The d-c power amplifier differs from the a-c power amplifier in several respects. First, transformers cannot be used for phase inversion at the input or for load matching at

the output. Second, both B^+ and B^- supplies must be employed if the output is to be balanced about ground. And third, if several stages of push-pull amplification are employed, the interstage coupling networks must pass direct current.

The electronic phase inverter has been discussed in Secs. 2.13 to 2.16. In that development both direct-coupled and a-c phase inverters were considered. Quite often, even in power amplifiers designed only for a-c service, phase inverters are employed in order to eliminate the weight and expense of the input transformer.

If the output of the amplifier is to be direct-coupled to the load, the output impedance of the amplifier should be made as low as possible in order to permit maximum power transfer. The output impedance of the amplifier may be made low by choosing tubes such as the 6AS7 that have

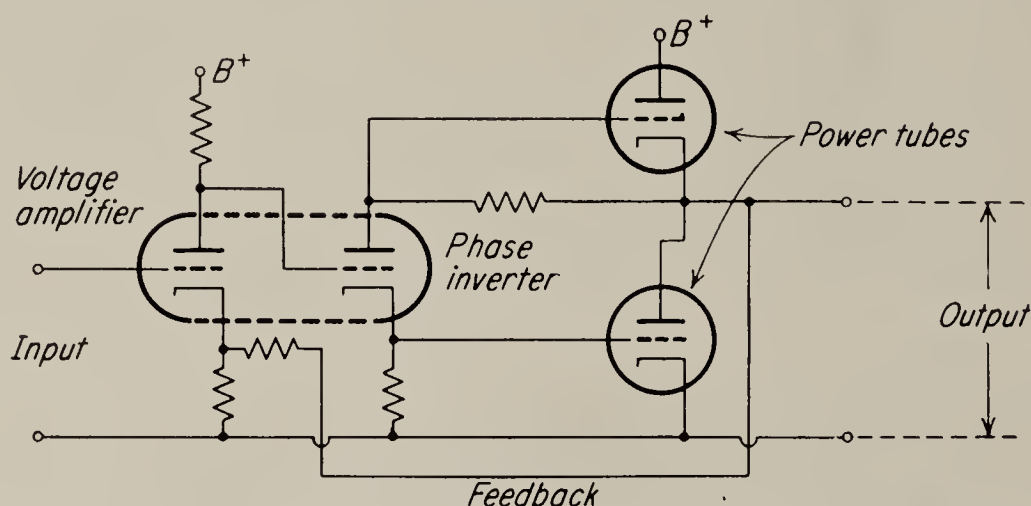


FIG. 3.11. Single-ended push-pull electronic amplifier with feedback.

low r_p . It is also common practice to employ a cathode-follower output stage to improve the load matching.

In addition to producing higher efficiency and lower distortion, push-pull operation of d-c amplifiers minimizes the effect of both plate- and filament-supply-voltage variation.

There are several interesting variations of the conventional electronic push-pull amplifier. Figure 3.11 shows a direct-coupled version of the so-called single-ended push-pull amplifier.¹ The input half of the double triode is connected as an amplifier with unbypassed cathode resistance. This acts as negative feedback for the input stage and also provides an input for feedback around the whole amplifier. The second half of the double triode is a phase inverter of the cathode-follower type discussed in Sec. 2.14. The phase inverter is direct-coupled to the power-output tubes. The upper output tube is driven from grid to cathode, not grid to ground. The power-tube currents add in the load, and their grid voltages are in phase opposition; thus this is a true push-pull amplifier. Another

¹ A. Peterson, and D. Sinclair, A Single-ended Push-pull Audio Amplifier, *Proc. IRE*, vol. 40, p. 7, 1952.

way to look at the circuit is to consider the upper output tube as a cathode follower and the lower output tube as a variable cathode resistor.

In Fig. 3.11 one terminal of the output is at ground potential, which is the reason for calling this a single-ended amplifier. The one disadvantage of the circuit as shown in the figure is that the quiescent current in the load is not zero. By the addition of a B^- supply the quiescent load current can be made zero. In this case the lower load terminal will be returned to ground rather than to the cathode of the lower output tube.

It will be observed that the plate current of the phase-inverter stage must flow through the lower output tube; thus the design of the output stage and the phase inverter are interdependent. It will be found that the choice of tube types is restricted by the fact that the quiescent phase-inverter current establishes the operating points of the output tubes. In order to maintain balanced gain through both halves of the phase inverter, the cathode resistor and plate resistor must be kept essentially equal; thus the grid bias on the output tubes will be equal. However, because of the additional phase-inverter current carried by the lower output tube, its operating point will differ somewhat from that of the upper tube. The design procedure is as follows:

1. Choose a phase-inverter tube that will permit an adequate plate-current swing to handle the required signal amplitude. Set its approximate operating point by choosing its resistors and supply voltage.

2. Choose an output-tube type that will handle the required load swing with the plate voltage and grid bias established in 1.

3. Modify 1 in order to conform more closely to the requirements in 2. This may entail changing tube types as well as supply voltages and resistors.

4. The restrictions placed on the circuit are extreme, and it may not be possible to satisfy the operating requirements for all three tubes in an optimum manner. Then there are several additional possibilities. First, it is not necessary that the B^+ supply and the B^- supply be equal. Second, a small, 200- to 500-ohm variable resistor may be placed in the cathode lead of the lower output tube in order to provide an additional adjustment. In any case the design is essentially a trial-and-error process.

Push-pull transistor amplifiers can also be adapted for d-c applications. The transistor power amplifier shown in Fig. 3.7 can be modified for d-c operation by using a d-c phase inverter and operating the amplifier directly into a split-phase load without a matching transformer. For maximum power transfer the designer must have the input-impedance specification of the load at his direction. This solution of the problem is of the brute-force type and does not take advantage of the unique potentialities of the transistor. Full use of these potentialities is made in the *complementary* push-pull amplifier.

The principle of complementary operation was first discussed by Sziklai.¹ The Sziklai amplifier employs a matched pair of n - p - n and p - n - p junction transistors, as shown in Fig. 3.12. Since the transistors have characteristics each of which is the negative image of the other, they operate without the requirement of external phase reversal at the input and output.

In practically every sense this is the ideal push-pull amplifier. Both the input and output are single-ended and operate at ground potential.

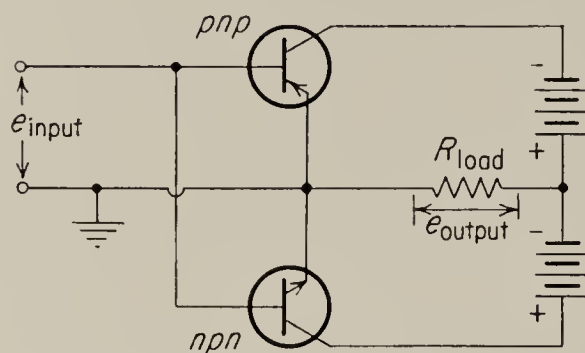


FIG. 3.12. The Sziklai amplifier employing complementary junction transistors.

No expensive and complicated coupling circuits are required, and the circuit is simplicity itself. Matched p - n - p , n - p - n junction pairs are essential to the successful operation of this circuit, and several such pairs are currently available. For instance, the 2N68-2N95 and the 2N101-2N102 pairs are available for amplifiers in the 10-watt range, and the list is soon to be extended. The conventional class B analysis holds here.

3.4. Thyatron Amplifiers. For many medium- and low-power applications thyatron control represents the most economical and lightweight solution. Thyatrons are essentially instantaneous in operation compared to the usual servo frequencies. Mercury-vapor thyatrons are somewhat sensitive to their environment, but inert-gas tubes have been developed that are as rugged as high-vacuum tubes.

The operation of a thyatron, or gas-filled tube, is fundamentally different from the operation of a vacuum tube. The gas tube acts as a relay or a switch rather than as a proportional device. Conduction cannot take place as long as the grid is held negative with respect to the cathode. If the grid-to-cathode voltage is gradually made less negative, the tube will suddenly begin conduction if the anode is positive with respect to the cathode. The grid-to-cathode potential at which conduction begins is a function of temperature and anode-to-cathode voltage. Figure 3.13 shows the grid potential for initiation of conduction plotted against positive anode potential for a typical thyatron.

Conduction is made up of electron flow from cathode to anode and by positive-ion flow from anode to cathode. Immediately after the start of conduction a sheath of positive ions builds up surrounding the grid and effectively insulates it from the remainder of the gas; thus the grid has no further effect on conduction² once the gas is ionized. The current flow is

¹ G. C. Sziklai, Symmetrical Properties of Transistors and Their Applications, *Proc. IRE*, vol. 41, p. 717, 1953.

² Dow, *op. cit.*, chap. 21.

not limited by the grid potential, and the conducting thyatron is effectively a short circuit from anode to cathode. If the anode current is not limited by the load, the tube will be damaged by the excessive current flow. There is a small voltage of 10 to 20 volts, depending on the gas, across the tube that is approximately independent of current, owing to the mechanism of gas conduction.¹ In order to stop conduction, the

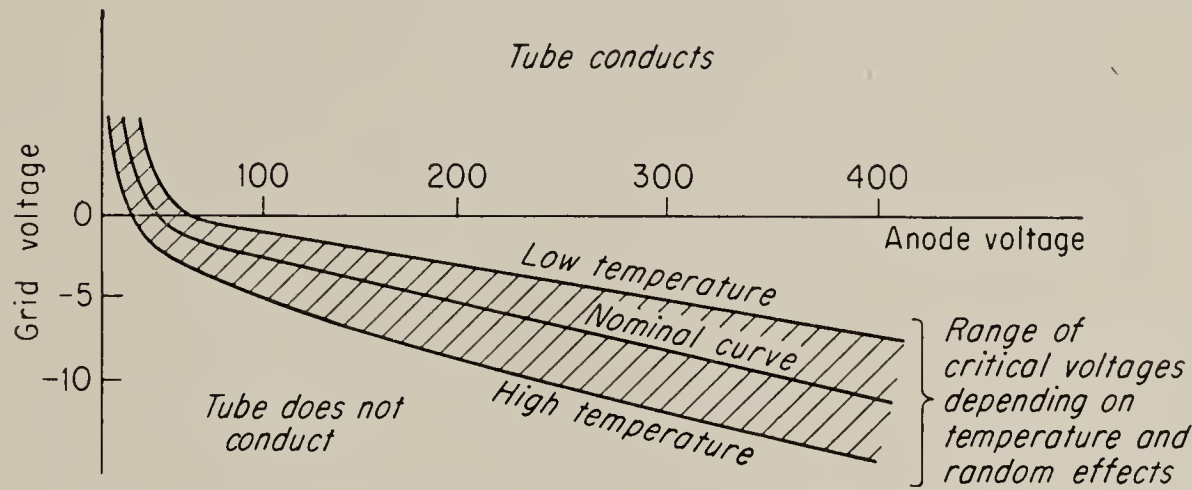


FIG. 3.13. Typical thyatron firing characteristics. Grid potential for initiation of conduction is plotted against positive anode potential for a typical thyatron.

positive anode voltage must be removed. This is usually accomplished by supplying the anode with an a-c anode voltage, thus operating the thyatron as a rectifier in which the average current is controlled by the grid potential.

3.5. Control of Firing Angle. Although there are a number of methods for controlling the point at which the thyatron begins to conduct, the

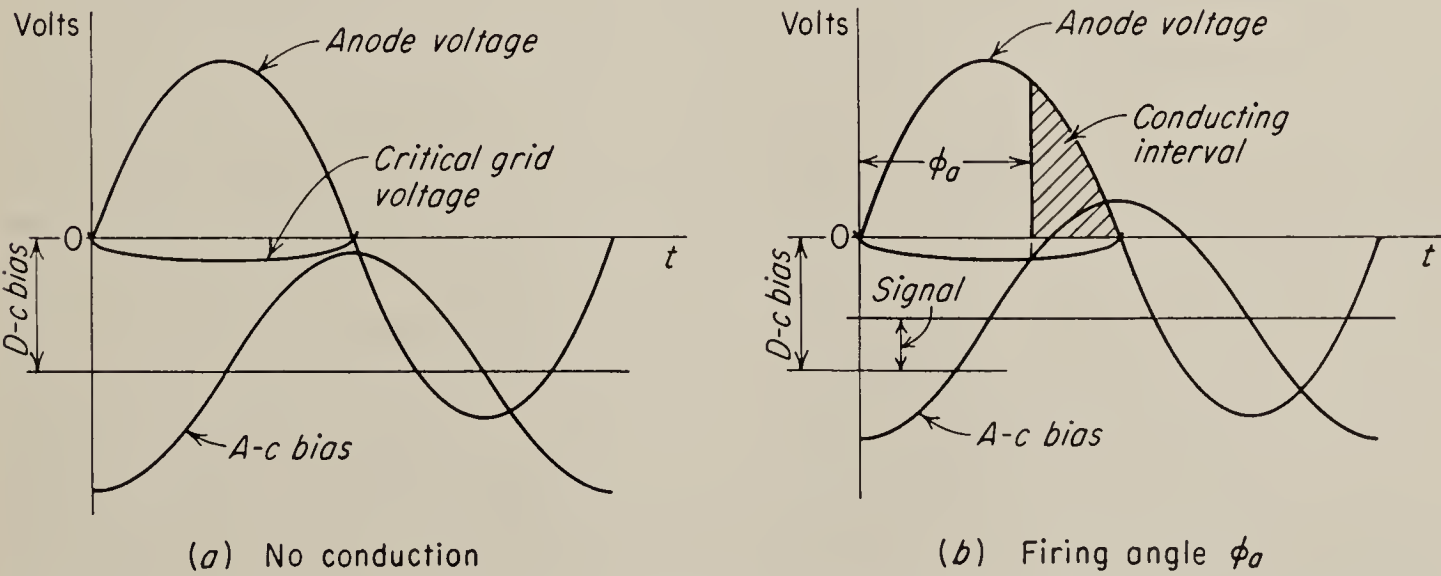


FIG. 3.14. Relation between anode and bias voltages for combination of a-c and d-c bias.

method most commonly used in control systems utilizes a combination of a-c and d-c grid bias together with a d-c signal voltage. As shown in Fig. 3.14, the a-c component of the grid bias consists of a voltage at line frequency, which lags the voltage applied to the anodes by 90°. The d-c

¹ *Ibid.*

component of the bias is approximately equal to the amplitude of the a-c component; hence with no signal the grid voltage just misses crossing the curve representing the critical value for which conduction starts. A positive signal applied in addition to the two bias voltages moves the total-grid-voltage wave up so that it now intersects the critical line at the angle ϕ_a shown in Fig. 3.14b. The angle ϕ_a is the firing angle, and the tube conducts during the interval shown shaded in Fig. 3.14b. Note that an increase in signal voltage results in a smooth advance of the firing angle. If the amplitude of the a-c component of the bias is very much larger than the critical grid potential, then the firing angle is approximately

$$\phi_a \approx \frac{\pi}{2} + \sin^{-1} \left(1 - \frac{E_i}{E_{dc}} \right) \quad (3.9)$$

where E_i is the signal voltage and E_{dc} is the d-c bias, which is assumed to be approximately equal to the amplitude of the a-c bias. For full control of the firing angle from 180° to 0° , E_i varies from zero to $2E_{dc}$. Thus the required signal range for full control is determined by the amplitude of the a-c bias. This can also be shown by computing the gain, or the change in firing angle degrees per volt of input signal, by differentiating Eq. (3.9) with respect to E_i . This gives

$$\frac{d\phi_a}{dE_i} = - \frac{1}{\sqrt{2E_iE_{dc} - E_i^2}} \quad (3.10)$$

If this is evaluated for $E_i = E_{dc}$, that is, for $\phi_a = 90^\circ$, the gain becomes

$$\left. \frac{d\phi_a}{dE_i} \right|_{\phi_a=90^\circ} = - \frac{1}{E_{dc}} \quad (3.11)$$

and is thus shown to be inversely proportional to the d-c bias. Thus the gain can be increased by decreasing the d-c bias and therefore the peak value of the a-c bias. Beyond a certain point, however, this reduction leads to difficulties. One of the important advantages of the control method using a-c bias is that the intersection of the grid voltage with the critical control line is sharp and not greatly affected by small variations of the critical voltage due to temperature or random effects. This advantage is largely lost if the amplitude of the a-c bias is reduced too greatly, and erratic firing may result.

The firing angle can also be controlled by shifting the phase of the a-c bias relative to the phase of the anode. In this method a fairly large a-c grid voltage is used, usually without d-c bias. The major disadvantage of this method is the difficulty of shifting the phase of the grid voltage. Although a number of circuits are available for this purpose,¹ they usually

¹ Reich, "Theory and Applications of Electron Tubes," McGraw-Hill Book Company, Inc., New York, 1944, pp. 510-515.

require a change of resistance or other circuit parameter. Hence the phase-shift method may occasionally be used when the control signal is in the form of a shaft position which can be used to vary an adjustable resistor. It is also possible to shift the phase electronically by the use of a reactance-tube circuit¹ or a saturable reactor. However, these methods are not so commonly used in servo systems as the control method using a-c bias described at the beginning of this section.

Extremely accurate control of the firing angle is possible by applying sharp voltage pulses to the grid.² In this method the signal is applied to a pulse-forming circuit in such a way that the time of the pulse is controlled by the signal. Although the firing angle can be regulated very precisely by this method, the circuitry required to produce the necessary pulses is usually too complicated to justify its use.

Direct application of the d-c signal to the grid without any a-c bias also permits a certain control of the firing angle because of the slight curvature of the critical-grid-voltage line. Despite its apparent simplicity, this method is not used in general, because it permits control of the firing angle only between 0 and 90°. Furthermore, since it depends on the curvature of the critical-grid-voltage line, it is very sensitive to any variations of this line, and firing tends to be erratic.

3.6. The Thyatron Amplifier with an Inductive Load. A simple thyatron control circuit which might be used to control the current flowing in the field of a d-c motor is shown in Fig. 3.15. The relative magnitudes

of the load inductance and resistance in this circuit have an important effect on the waveshape of the current that flows during periods of tube conduction. If the inductance is relatively small, the current and voltage waveshapes are identical. However, if the inductive reactance is appreciable, the build-up and decay of the current is delayed, and the current

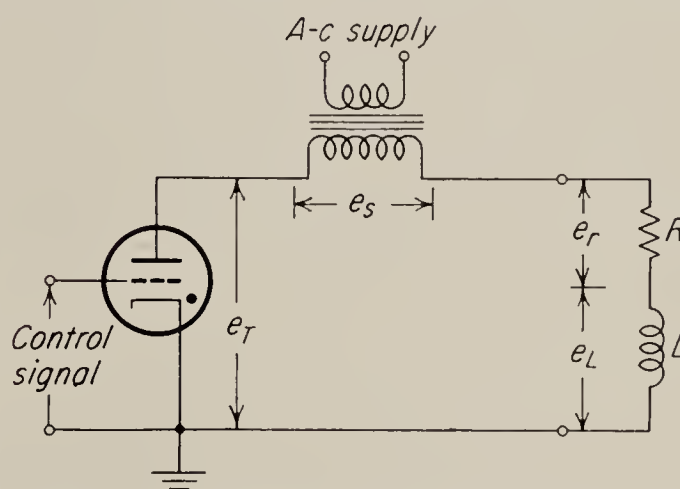


FIG. 3.15. Half-wave thyatron circuit with inductive load.

waveshape departs from that of the voltage. The determination of the average load current as a function of firing angle becomes a rather involved task under these conditions. Both graphical and analytical techniques are used. The graphical method has the advantage of displaying the fundamental processes taking place in the circuit somewhat more clearly, but when accurate numerical results are required, the analytical method is usually preferable.

¹ *Ibid.*

² J. G. Skalnik, Pulse Control Thyatron, *Electronics*, December, 1949, p. 120.

Typical voltage and current waveshapes found in the half-wave circuit of Fig. 3.15 are shown in Fig. 3.16. The components of voltage are labeled in Fig. 3.15. At point a the tube ionizes, and e_T drops to a very small value that is essentially independent of anode current. The remainder of the voltage e_s divides across R and L . The voltage e_L

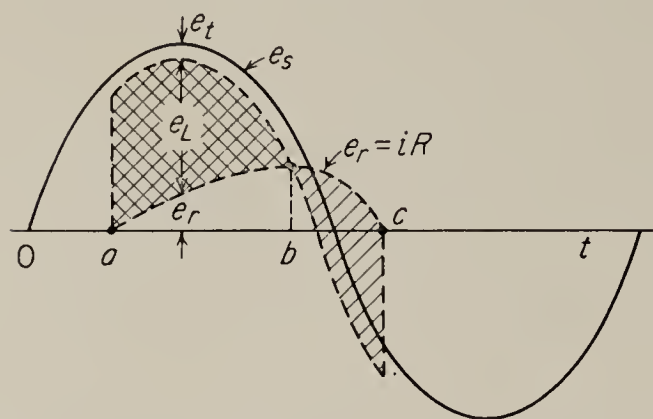


FIG. 3.16. Voltage and current wave-shapes for circuit in Fig. 3.15.

is equal to $L di/dt$, and e_r is equal to Ri . The voltages e_L and e_r may be found by point-by-point calculations. Calculations should be made in a logical sequence, e.g., that given by Chin and Moyer.¹ At point a the current is zero; therefore iR is zero, and

$$e_L = e_s - e_T \quad (3.12)$$

where e_T is a constant, usually taken as 15 volts or perhaps neglected.

Having found e_L at point (a) , we write

$$\frac{di}{dt} = \frac{e_L}{L} \quad (3.13)$$

thus establishing the slope of the current wave. The initial slope of e_r is found by multiplying the initial value of di/dt by R . The increment over which this initial slope is extended depends on the accuracy required for the results. If it is sufficient to find the initial slope, the maximum value, and the point at which e_r once more becomes zero, the calculation is fairly simple. We know that the maximum value (or zero slope) of the iR voltage wave falls at point b , since at that point e_L and therefore di/dt are zero, and

$$e_r = e_s - e_T \quad (3.14)$$

As yet, we have not determined point b . We pick a point b such that the integral of e_L from a to b times R/L is equal to the value of e_r at point b . This may be seen by noting that

$$e_L = L \frac{di}{dt} \quad (3.15)$$

$$\text{Therefore} \quad \int_a^b e_L dt = \int_{i_a=0}^{i_b} L di \quad (3.16)$$

$$\text{and} \quad \int_a^b e_L dt = Li_b \quad (3.17)$$

$$\text{Now} \quad i_b = \frac{e_{rb}}{R} \quad (3.18)$$

¹ P. T. Chin and E. E. Moyer, Controlled Rectifier Circuits, *Trans. AIEE*, vol. 63, p. 501, 1944.

Substituting (3.18) into (3.17), we have

$$e_{rb} = \frac{R}{L} \int_a^b e_L dt \quad (3.19)$$

The integral is proportional to the area shown crosshatched in Fig. 3.16.

The argument used to obtain the build-up of current to point b also holds during the decay of the current. The point c at which i becomes zero, the so-called extinction point, may be found in the same manner as point b . A point c is chosen such that the shaded area between points b and c is equal to the crosshatched area between points a and b found above. Since an inductor must return as much energy to the line as it receives and since the two areas represent positive and negative energy, respectively, we see that the areas must be equal.

In general, analytic methods are preferred over graphical methods. We shall apply analytic methods to the same circuits as an example. Note that the voltage applied to the load is

$$e(t) = E_s \sin(\omega t + \phi_a) - E_T \quad (3.20)$$

where E_s is the amplitude of the supply voltage, E_T the tube drop, ω the line frequency in radians per second, and ϕ_a the firing angle. The time origin ($t = 0$) is the time at which the tube fires.

The instantaneous load current is best found as a function of time by use of Laplace transform methods. The Laplace transform of the voltage of Eq. (3.20) is

$$\hat{E} = \frac{E_s(\omega \cos \phi_a + s \sin \phi_a)}{s^2 + \omega^2} - \frac{E_T}{s} \quad (3.21)$$

The impedance of the circuit in terms of the complex variable s is

$$\hat{Z} = R + Ls = L \left(s + \frac{R}{L} \right) \quad (3.22)$$

Therefore the transform of the load current is

$$\hat{I} = \frac{E_s}{L} \frac{\omega \cos \phi_a + s \sin \phi_a}{(s^2 + \omega^2)(s + R/L)} - \frac{E_T}{L} \frac{1}{s(s + R/L)} \quad (3.23)$$

To find the current as a function of time, we perform an inverse transformation on this expression. This results in

$$i(t) = \frac{E_s}{Z} \left[\sin(\omega t + \phi_a - \theta) - \sin(\phi_a - \theta) \exp\left(-\frac{R}{L} t\right) \right] - \frac{E_T}{R} \left[1 - \exp\left(-\frac{R}{L} t\right) \right] \quad (3.24)$$

In this equation $Z = \sqrt{R^2 + \omega^2 L^2}$ and $\theta = \tan^{-1} \omega L/R$. Although this

expression can be used to find the instantaneous value of the load current at any time t , it cannot be solved directly for the time at which the current goes through zero, i.e., the extinction point, because of the simultaneous presence of exponential and trigonometric functions of time. It is, however, possible to obtain the same information indirectly by assuming an arbitrary value of ωt and finding the value of firing angle ϕ_a required to make the current go through zero at this point. If $i(t)$ in Eq. (3.24) is set equal to zero, the equation becomes

$$\sin(\omega t + \phi_a - \theta) - \sin(\phi_a - \theta) \exp\left(-\frac{R}{L}t\right) = \frac{E_T Z}{E_s R} \left[1 - \exp\left(-\frac{R}{L}t\right)\right] \quad (3.25)$$

This equation may be rearranged to give

$$\begin{aligned} \phi_a = \theta + \pi \\ - \sin^{-1} \left[\frac{E_T Z [1 - \exp(-(R/L)t)]}{E_s R \sqrt{1 - 2 [\exp(-(R/L)t)] \cos \omega t + \exp(-2(R/L)t)}} \right] \\ - \tan^{-1} \left[\frac{\sin \omega t}{\cos \omega t - \exp(-(R/L)t)} \right] \end{aligned} \quad (3.26)$$

Some care must be used in evaluating the inverse trigonometric functions in this expression, since these functions are multivalued. If the result is examined carefully, however, it is clear that the \sin^{-1} function represents angles from 0 to 90° while the \tan^{-1} function represents angles from 0 to 180° . The time t in Eq. (3.26) is the time at which the tube conduction ceases, measured from the firing instant. Hence if ϕ_c is the extinction angle, we have

$$\phi_c - \phi_a = \omega t \quad (3.27)$$

where both ϕ_a and ϕ_c are measured from point O in Fig. 3.16. Substituting for ωt in Eq. (3.26), we obtain

$$\begin{aligned} \phi_a = \theta + \pi - \sin^{-1} \\ \frac{E_T Z \left\{ 1 - \exp \left[-\frac{R}{\omega L} (\phi_c - \phi_a) \right] \right\}}{E_s R \sqrt{1 - 2 \exp \left[-\frac{R}{\omega L} (\phi_c - \phi_a) \right] \cos (\phi_c - \phi_a) + \exp \left[-\frac{2R}{\omega L} (\phi_c - \phi_a) \right]}} \\ - \tan^{-1} \frac{\sin (\phi_c - \phi_a)}{\cos (\phi_c - \phi_a) - \exp \left[-\frac{R}{\omega L} (\phi_c - \phi_a) \right]} \end{aligned} \quad (3.28)$$

Thus we find that the firing angle ϕ_a can be determined analytically in terms of the difference between extinction and firing angles. Note that a considerable simplification results if the tube drop E_T is neglected, since an entire term of Eq. (3.28) will then disappear. Also note that, if the

term $(R/\omega L)(\phi_c - \phi_a)$ is large (greater than about 3), all the exponential terms become negligibly small. Then, approximately,

$$\phi_a = \theta + \pi - \sin^{-1} \frac{E_T Z}{E_s R} - \phi_c + \phi_a$$

or

$$\phi_c = \theta + \pi - \sin^{-1} \frac{E_T Z}{E_s R} \quad (3.29)$$

This indicates that, if the inductance is sufficiently small and if the difference between extinction and firing angles is sufficiently large, the extinction point becomes independent of the firing point. A typical set of

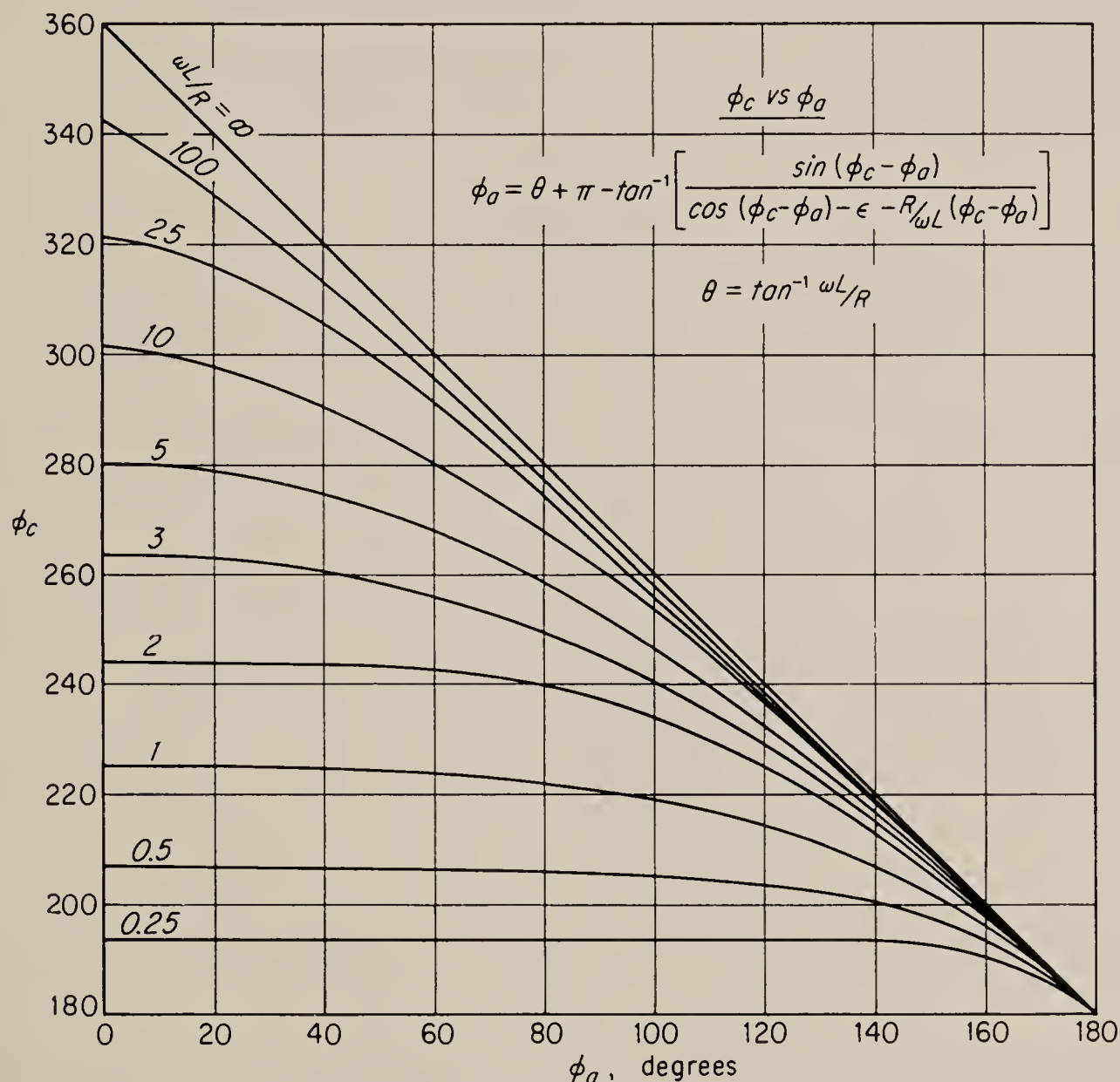


FIG. 3.17. Relation between firing angle and extinction angle for half-wave circuit with inductive load. Tube drop is neglected.

curves of ϕ_a versus ϕ_c for various values of $\omega L/R$ is given in Fig. 3.17. The tube drop has been neglected in drawing these curves.

Once the extinction point has been found, the determination of the average load current becomes relatively simple. Although it is possible to find the average current directly by integrating Eq. (3.24) over the proper limits, it is easier first to find the average voltage and then to divide by the load resistance. For the half-wave circuit considered here,

the average voltage applied to the load is

$$\begin{aligned} E_{av} &= \frac{1}{2\pi} \int_{\phi_a}^{\phi_c} (E_s \sin \phi - E_T) d\phi \\ &= \frac{1}{2\pi} [E_s (\cos \phi_a - \cos \phi_c) - E_T(\phi_c - \phi_a)] \end{aligned} \quad (3.30)$$

The average load current is now simply E_{av}/R or

$$I_{av} = \frac{1}{2\pi R} [E_s (\cos \phi_a - \cos \phi_c) - E_T(\phi_c - \phi_a)] \quad (3.31)$$

If the inductance is sufficiently small and if the firing point comes sufficiently early so that $(R/\omega L)(\phi_c - \phi_a) > 3$, the extinction angle may be found from Eq. (3.29). Under these conditions the average current becomes

$$\begin{aligned} I_{av} &= \frac{1}{2\pi R} \left[E_s \cos \phi_a + \frac{1}{Z} \sqrt{E_s^2 R^2 - E_T^2 Z^2} \right. \\ &\quad \left. + E_T \left(\frac{X}{R} + \cos \phi_a - \tan^{-1} \frac{X}{R} - \pi + \sin^{-1} \frac{E_T Z}{E_s R} \right) \right] \end{aligned} \quad (3.32)$$

where $X = \omega L$, $Z = (R^2 + X^2)^{1/2}$, and where we have used the fact that $\theta = \tan^{-1} \omega L/R$. If the tube drop E_T is negligible, this becomes

$$I_{av} = \frac{E_s}{2\pi} \left(\frac{\cos \phi_a}{R} + \frac{1}{Z} \right) \quad (3.33)$$

Here it is clear that an increase of inductance, which increases Z , will result in less average current.

When the inductance is very large, so that $R/\omega L \approx 0$, Eq. (3.28) may be solved approximately to give

$$\phi_c \approx \pi + 2\theta - \phi_a$$

But if the inductance is very large, $\theta \approx \pi/2$. Hence approximately

$$\phi_c \approx 2\pi - \phi_a \quad (3.34)$$

This result implies that for very large inductance the conduction period for negative values of instantaneous supply voltage is approximately equal to the conduction period for positive voltage; i.e., the shaded area in Fig. 3.16 is equal to the crosshatched area. The average current is therefore approximately equal to zero. This is borne out analytically by substituting Eq. (3.34) into Eq. (3.31). If the tube drop is neglected, the result will be seen to be zero. In practice the load current is, of course, not zero, since Eq. (3.34) is only approximate; however the effect of large inductance in severely reducing the average load current should be evident.

Curves of average current as a function of firing angle for several values of $\omega L/R$ are given in Fig. 3.18. These curves are obtained as follows: First, a number of values of $\phi_c - \phi_a$ are chosen and used in Eq. (3.28) to determine ϕ_a . Then ϕ_c can also be found, and the current is then obtained from Eq. (3.31). The tube drop has been neglected, and the

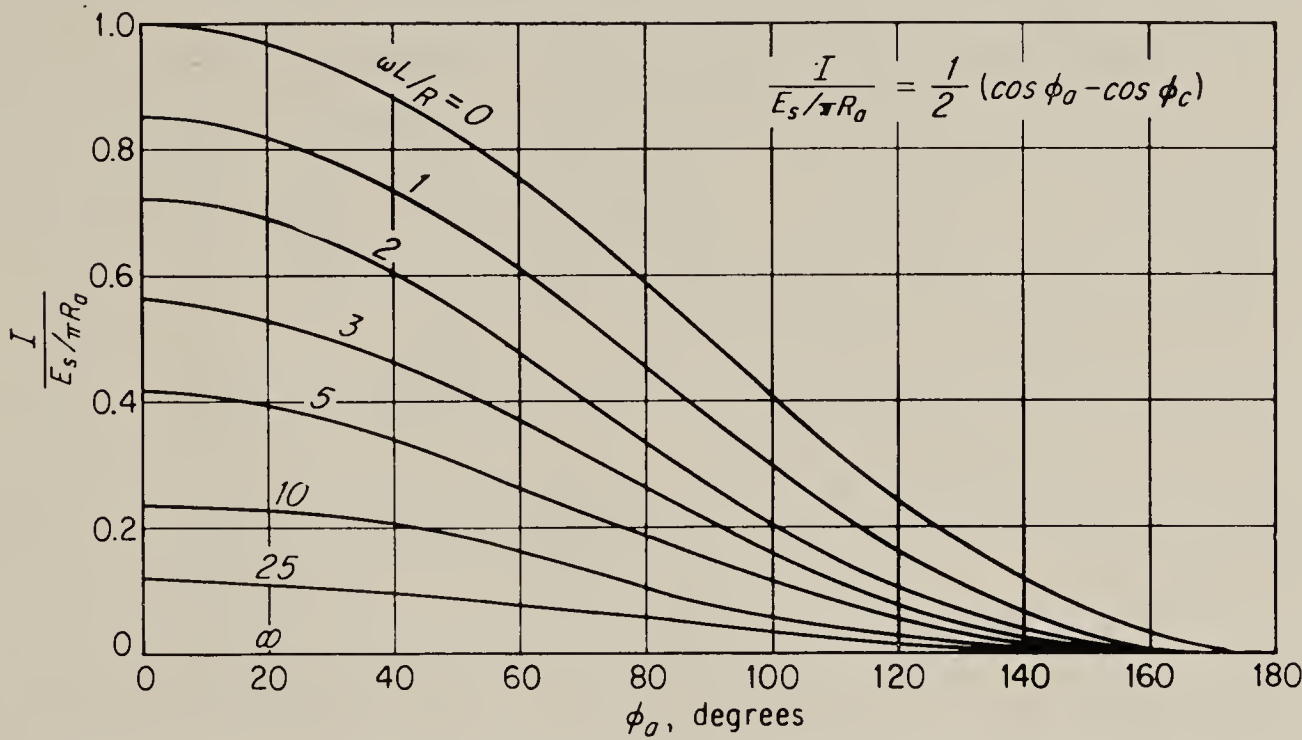


FIG. 3.18. Normalized average load current versus firing angle for half-wave thyatron circuit. Tube drop is neglected.

curves have been normalized with respect to the maximum current in a purely resistive load, $E_s/\pi R_a$.

In order to increase the load current in half-wave circuits driving highly inductive loads, use is sometimes made of the so-called *freewheeling circuit*, shown in Fig. 3.19a. In this circuit the diode across the load

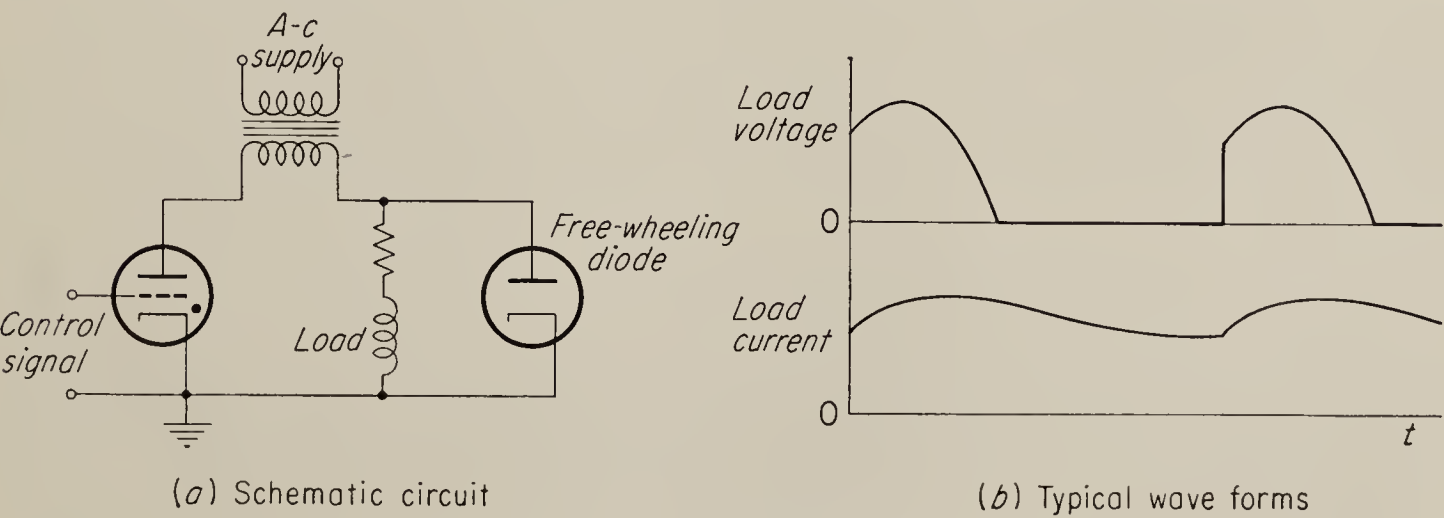


FIG. 3.19. Freewheeling circuit.

conducts during the negative half cycle of a-c supply voltage and provides a low-resistance discharge path for the load current. Thus the load current is not forced down by the negative supply voltage, as is the case in the half-wave thyatron circuit without a diode, but instead it follows a more slowly decreasing, exponential discharge curve. This is shown in

Fig. 3.19b, which shows both the voltage and current waveshapes. If the tube drop of the diode is negligible, the load voltage during negative half cycles of a-c supply voltage is zero, and the waveshape is identical with that observed on a half-wave circuit with a pure resistance load. Hence the average load current is equal to the average current found for pure resistance, $(E_s/2\pi R)(1 + \cos \phi_a)$.

3.7. The Full-wave Thyatron Amplifier with Inductive Load. The full-wave circuit shown in Fig. 3.20 is more commonly used than the half-wave circuit considered in Sec. 3.6. Since there are two voltage pulses per cycle of supply voltage, a smoother current flow to the load results. The analysis of the full-wave circuit is complicated somewhat by the fact that under certain conditions the load current may be continuous. This happens, in fact,

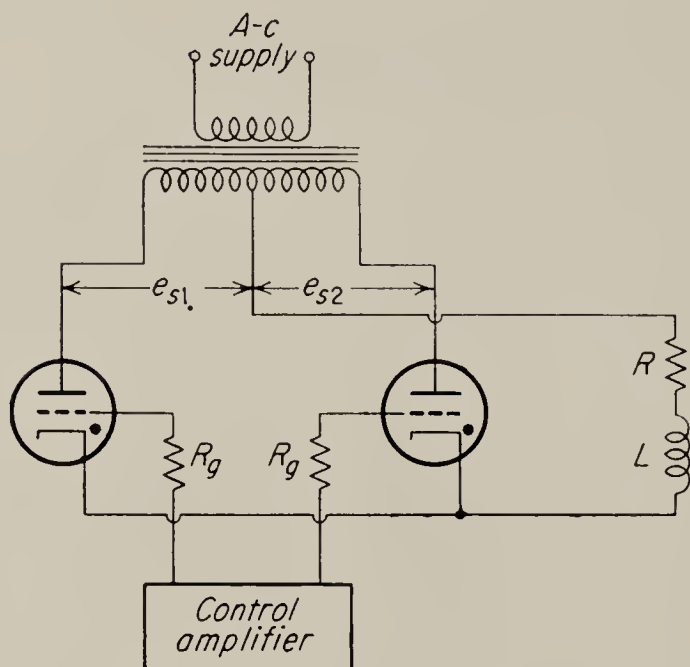


FIG. 3.20. Full-wave thyatron circuit. R_g limits the current drawn by the control grid.

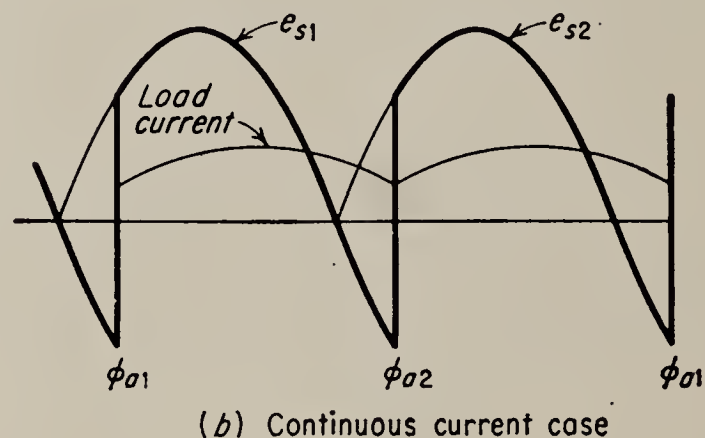
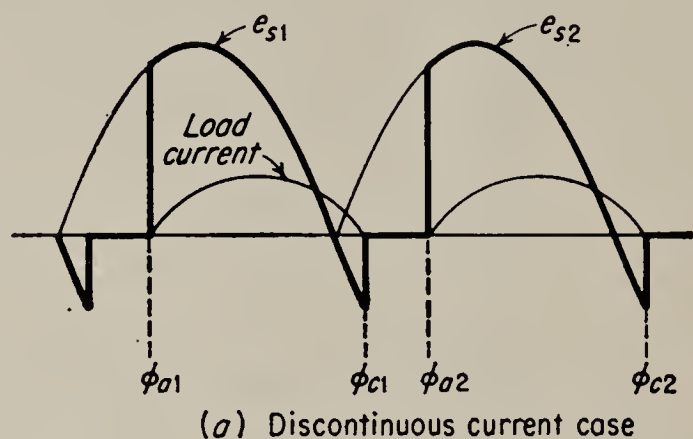


FIG. 3.21. Voltage and current waveshapes in full-wave thyatron circuit.

whenever the extinction angle of one tube as given by Eq. (3.28) occurs later than the firing point of the other tube. Typical waveshapes of voltage and current, for both the continuous and discontinuous case, are shown in Fig. 3.21. Since in normal operation the firing points of the two tubes are 180° apart, the condition for continuous current flow is

$$\phi_c - \phi_a \geq \pi$$

where both ϕ_c and ϕ_a are measured with respect to the same origin and on the same tube. Substituting this relation into Eq. (3.28) gives

$$\phi_a \leq \theta - \sin^{-1} \frac{E_T Z}{E_s R} \frac{1 - \exp(-R\pi/\omega L)}{1 + \exp(-R\pi/\omega L)} \quad (3.35)$$

for the condition for continuous current flow. In making this substitution, it should be noted that the last term of Eq. (3.28) must have the value π rather than zero.

The determination of average voltage and current is now reasonably straightforward. If the circuit operates in the discontinuous mode, the voltage and current are found exactly as for the half-wave circuit except that the averages must be taken over one half cycle. Thus

$$\begin{aligned} E_{av} &= \frac{1}{\pi} \int_{\phi_a}^{\phi_c} (E_s \sin \phi - E_T) d\phi \\ &= \frac{1}{\pi} [E_s (\cos \phi_a - \cos \phi_c) - E_T(\phi_c - \phi_a)] \end{aligned} \quad (3.36)$$

while the average load current is

$$I_{av} = \frac{1}{\pi R} [E_s (\cos \phi_a - \cos \phi_c) - E_T(\phi_c - \phi_a)] \quad (3.37)$$

If the inductance is small and if the firing point comes early enough, expressions similar to Eqs. (3.32) and (3.33) can be derived for the full-wave circuit; in fact the current is simply twice the value given in these equations.

When the circuit operates in the continuous mode, the average voltage and current are obtained more easily, since it is only necessary to integrate from one firing point to the next. This is made clear by inspection of Fig. 3.21b. Thus, for the continuous mode the average voltage is given by

$$\begin{aligned} E_{av} &= \frac{1}{\pi} \int_{\phi_a}^{\phi_a + \pi} (E_s \sin \phi - E_T) d\phi \\ &= \frac{2}{\pi} E_s \cos \phi_a - E_T \end{aligned} \quad (3.38)$$

and the average current becomes

$$I_{av} = \frac{2}{\pi R} E_s \cos \phi_a - \frac{E_T}{R} \quad (3.39)$$

Although the inductance no longer appears explicitly in Eq. (3.39), the load current is still affected by it, since the inductance is responsible for the continuous conduction of the tubes. This causes negative voltage to be applied to the load for part of the cycle and results in a reduction of average current (see Fig. 3.21). The rate of increase of current with firing angle is, however, more rapid in the continuous mode than in the discontinuous mode. This is indicated by the sharp change in slope of the curve of current versus firing angle. Typical curves of this type are shown in Fig. 3.22.

3.8. Thyatron-amplifier Transfer Function. Up to this point only the static performance of the thyatron amplifier has been considered. In control applications one is, however, also very much interested in the transient performance or, in general, in something equivalent to a transfer function. The transient behavior of the circuit depends on whether it is operating in the discontinuous or the continuous mode. If the circuit operates in the discontinuous mode, each current pulse represents an entirely new transient, which is completely independent of the previous current pulse. Thus in the discontinuous mode the thyatron amplifier may be characterized by a *one-cycle response time*; i.e., the current reaches a new average value within one cycle of the change of firing angle.

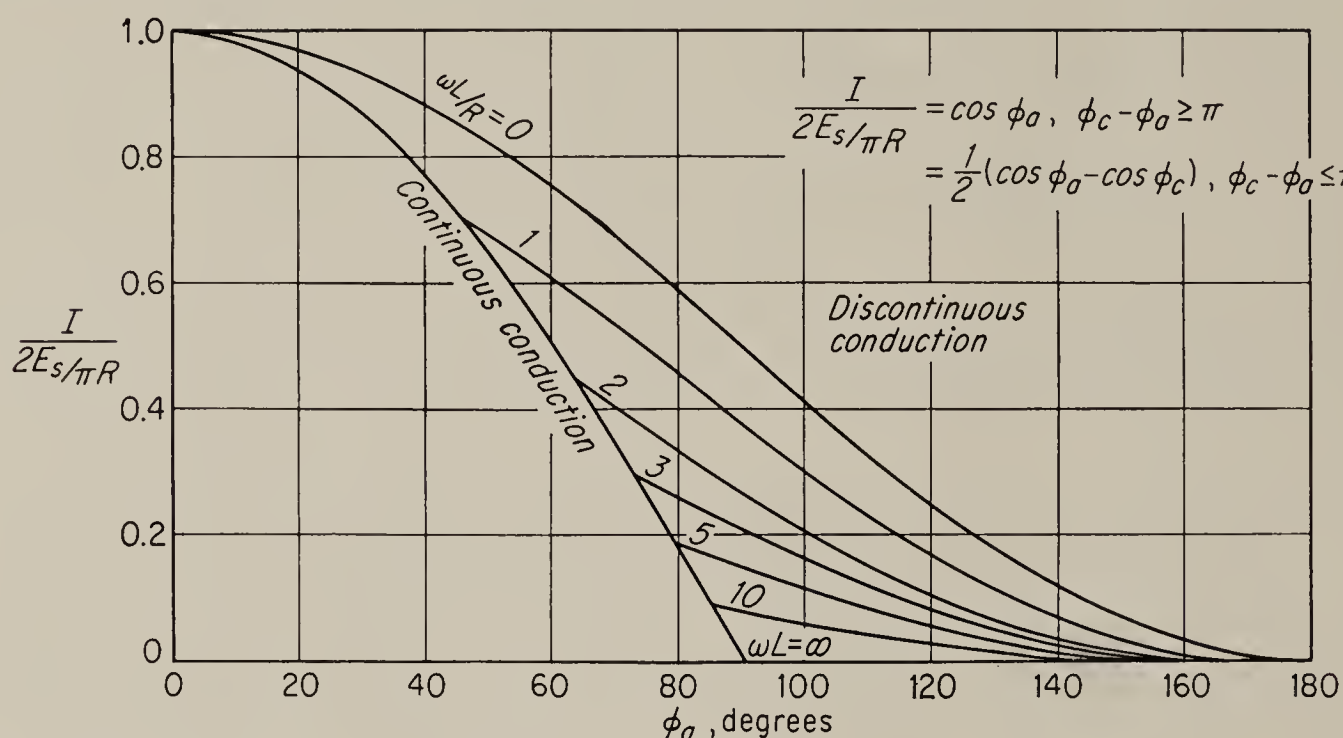


FIG. 3.22. Variation of average load current with firing angle in full-wave circuit.

Since the current in the half-wave circuit is always discontinuous if no freewheeling diode is used, this circuit always has a *half-cycle response time*. However, when the full-wave circuit enters the continuous mode, its behavior changes. The average voltage still has a one-cycle response, but the current now increases exponentially with the time constant L/R . It should be noted that, since the thyatron amplifier is a nonlinear circuit, a transfer function in the usual sense does not exist. However, in the continuous case one may define a “quasi-linear” transfer function, which relates small variations of load current to small variations of firing angle. Such a transfer function has the form

$$\frac{\Delta \hat{I}}{\Delta \hat{\phi}_a} = \frac{k}{(L/R)s + 1} \quad (3.40)$$

In this equation k is the slope of the curve of current versus firing angle and is a function of the operating point around which the small changes ΔI and $\Delta \phi_a$ are taken.

3.9. The Thyatron Amplifier with a D-C-motor Load. Thyatron amplifiers are quite often used to control low- and medium-power d-c motors. A simple half-wave circuit for unidirectional speed control is shown in Fig. 3.23 with all voltages and currents labeled. The motor field is assumed to be excited by a separate d-c supply and is assumed to be fixed.

A d-c-motor load differs from loads considered up to now in that it generates a counter emf $E_g = k_v \Omega$, where Ω is the speed and k_v the velocity voltage coefficient (see Sec. 4.15). When the speed is constant, this counter emf is a fixed voltage and may be represented by a battery in the load circuit. As a result of the presence of the

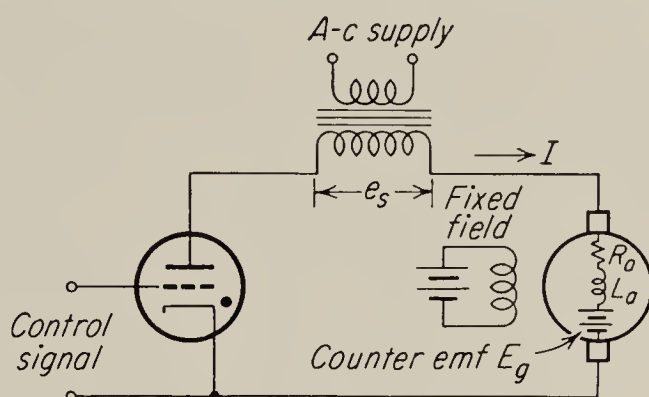


FIG. 3.23. Half-wave thyatron circuit with motor load.

counter emf, the thyatron may not fire when the grid potential passes the critical value, and it is convenient, therefore, to define three modes of operation for the circuit. These modes are defined by reference to Fig. 3.24.

When the grid passes the critical value in region 1 of Fig. 3.24, the tube cannot conduct, since the voltage on its plate is negative. However, unless the grid signal consists of short pulses, the tube will conduct as

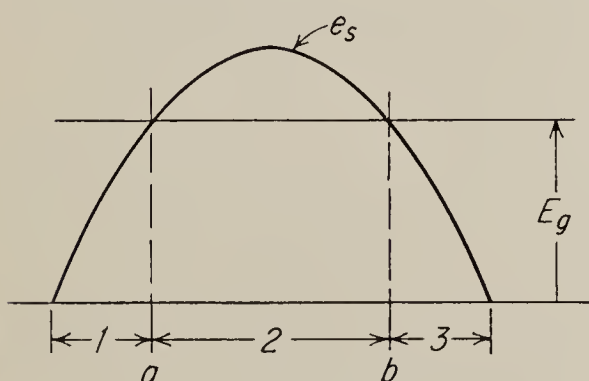


FIG. 3.24. Operating modes of thyatron amplifier with motor load.

soon as the instantaneous a-c supply voltage exceeds the counter emf E_g (point a in Fig. 3.24). Thus in mode 1 changes in firing angle that do not go beyond point a have no effect on the current flowing in the motor. A firing angle occurring anywhere between points a and b in Fig. 3.24 puts the circuit into operating mode 2. In this mode the circuit operates normally, with the tube firing as soon as the critical grid potential

is exceeded. Mode 3 is defined by a firing point occurring in region 3. The armature current is always zero in this mode. The three modes are defined mathematically by the following three equations:

$$\begin{aligned}
 \text{For mode 1:} \quad & 0 \leq \phi_a \leq \sin^{-1} \frac{E_g}{E_s} \\
 \text{For mode 2:} \quad & \sin^{-1} \frac{E_g}{E_s} \leq \phi_a \leq \pi - \sin^{-1} \frac{E_g}{E_s} \\
 \text{For mode 3:} \quad & \pi \geq \phi_a \geq \pi - \sin^{-1} \frac{E_g}{E_s}
 \end{aligned} \tag{3.41}$$

where ϕ_a is the firing angle. Note that the \sin^{-1} functions used in these equations are the so-called principal values. Note also that, for a given value of ϕ_a , the circuit can be in only one of two modes. Thus for $\phi_a < \pi/2$ the circuit goes from mode 2 to mode 1 as E_g is increased. For $\phi_a > \pi/2$ the circuit goes from mode 2 to mode 3 as E_g is increased.

The analysis of the operation of the motor is simplified considerably if it is assumed that the inductance of the armature circuit is negligible. This is true for most d-c motors, particularly those designed for use with thyratrons, because, as will be shown later, the presence of inductance reduces the maximum current and developed torque. Thus, in the following analysis we assume that the inductance is zero. Also, it is convenient, although not necessary, to neglect the tube drop E_T .

If a firing angle and the speed of the motor are given, the average armature current can be found as follows: If the circuit is in mode 1, the tube conducts during the time interval bounded by points a and b in Fig. 3.24. Hence the average current is

$$\begin{aligned} I &= \frac{1}{2\pi R_a} \int_{\phi_a}^{\phi_b} (E_s \sin \phi - E_g) d\phi \\ &= \frac{1}{2\pi R_a} [E_s (\cos \phi_a - \cos \phi_b) - E_g (\phi_b - \phi_a)] \end{aligned} \quad (3.42)$$

where R_a is the armature resistance. But ϕ_a and ϕ_b are defined by

$$\begin{aligned} \phi_a &= \sin^{-1} \frac{E_g}{E_s} \\ \phi_b &= \pi - \sin^{-1} \frac{E_g}{E_s} \end{aligned} \quad (3.43)$$

$$\text{Hence} \quad I = \frac{E_s}{2\pi R_a} \left[2 \sqrt{1 - \left(\frac{E_g}{E_s} \right)^2} - \frac{E_g}{E_s} \left(\pi - 2 \sin^{-1} \frac{E_g}{E_s} \right) \right] \quad (3.44)$$

If the circuit is in mode 2, the thyatron fires at the firing point ϕ_a , which is now no longer determined by E_g but may have any arbitrary value. Hence

$$I = \frac{1}{2\pi R_a} \int_{\phi_a}^{\pi - \sin^{-1}(E_g/E_s)} (E_s \sin \theta - E_g) d\theta \quad (3.45)$$

$$= \frac{E_s}{2\pi R_a} \left[\cos \phi_a + \sqrt{1 - \frac{E_g^2}{E_s^2}} + \frac{E_g}{E_s} (\phi_a - \pi) + \frac{E_g}{E_s} \sin^{-1} \frac{E_g}{E_s} \right] \quad (3.46)$$

Since the developed torque of a motor is given by $Q_d = k_t I$ [see Eq. (4.56)] and since $E_g = k_v \Omega$, Eqs. (3.45) and (3.46) may be rewritten in terms of torque, speed, and firing angle as follows:

For mode 1:

$$Q_d = \frac{k_t E_s}{2\pi R_a} \left[2 \sqrt{1 - \left(\frac{k_v \Omega}{E_s} \right)^2} - \frac{k_v \Omega}{E_s} \left(\pi - 2 \sin \frac{k_v \Omega}{E_s} \right) \right] \quad (3.47)$$

For mode 2:

$$Q_d = \frac{k_t E_s}{2\pi R_a} \left[\cos \phi_a + \sqrt{1 - \left(\frac{k_v \Omega}{E_s} \right)^2} + \frac{k_v \Omega}{E_s} (\phi_a - \pi) + \frac{k_v \Omega}{E_s} \sin^{-1} \frac{k_v \Omega}{E_s} \right] \quad (3.48)$$

These two equations can be used to plot curves of developed torque versus speed for different firing angles. Such a set of curves is shown in Fig. 3.25. These curves have been normalized with respect to maximum

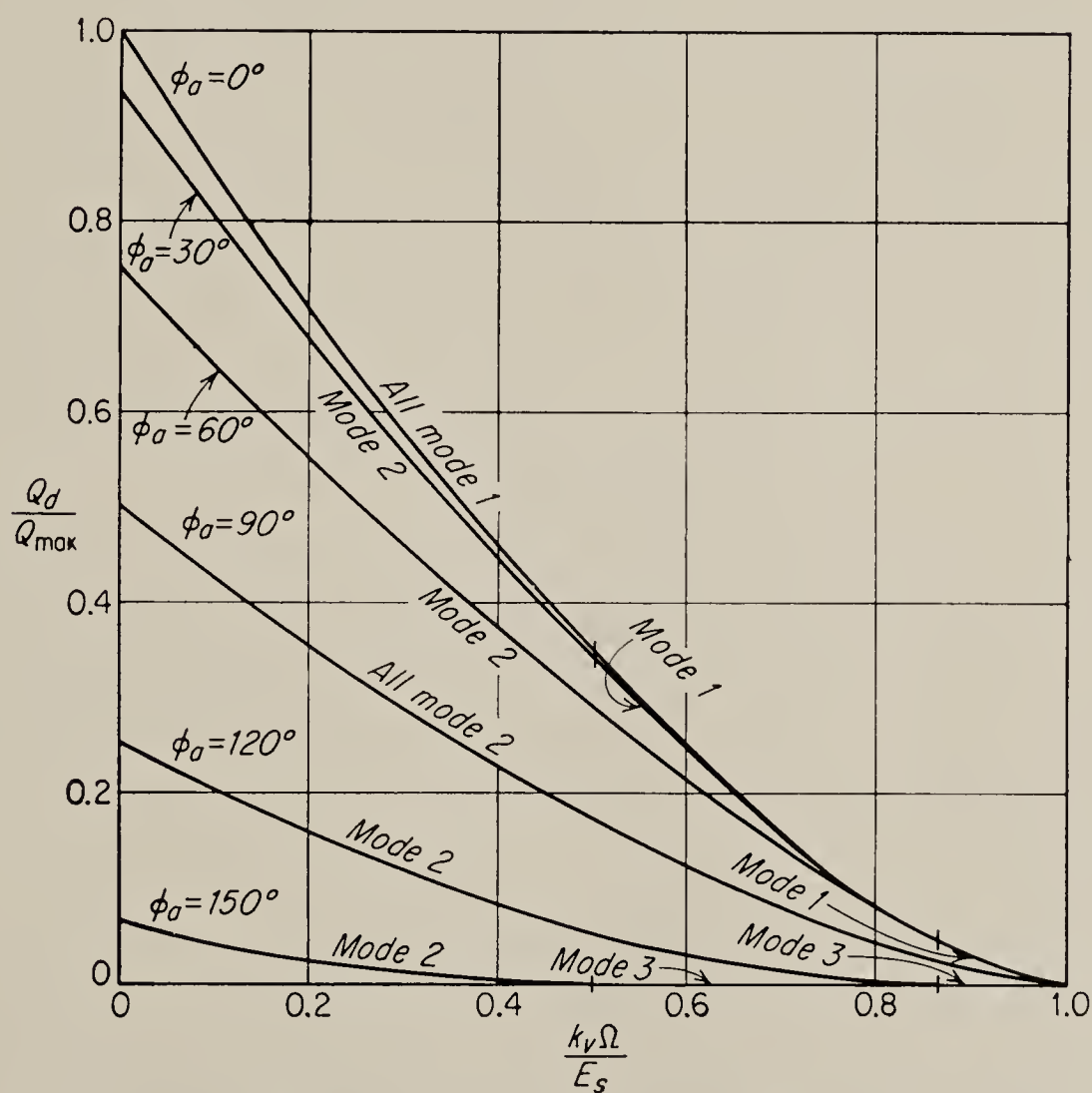


FIG. 3.25. Speed-torque curves for thyatron-controlled motor.

torque and speed of the motor. The points of transition between modes are also indicated in this figure. Note that the transition from mode 1 to mode 2 is not abrupt; in fact it can be shown by differentiating Eqs. (3.47) and (3.48) that the slopes of the two curves are the same at the transition point. The same is true for the transition from mode 2 to mode 3.

A number of interesting points can be deduced from Fig. 3.25. Note that, for firing angles less than 90° and for small developed torque, the motor operates almost entirely in mode 1. Its speed is therefore not con-

trolled by the firing angle; in fact it appears from the curves that the no-load speed for all firing angles less than 90° is equal to the maximum speed. If the no-load speed of a motor is to be controlled, firing angles of more than 90° must be employed. At these firing angles the circuit operates, however, in a highly nonlinear on-off region, i.e., the boundary between mode 2 and mode 3. Under certain conditions this may give rise to a nonlinear type of oscillation. Thus, consider the circuit shown in Fig. 3.26. Suppose that for some reason the motor is running too fast. The feedback circuit acts to retard the firing angle. If the motor is running at no load, the circuit goes into mode 3, and the armature current is cut off. The motor now coasts until its speed is less than the reference speed. If there is a relatively large time lag in the controller, the motor

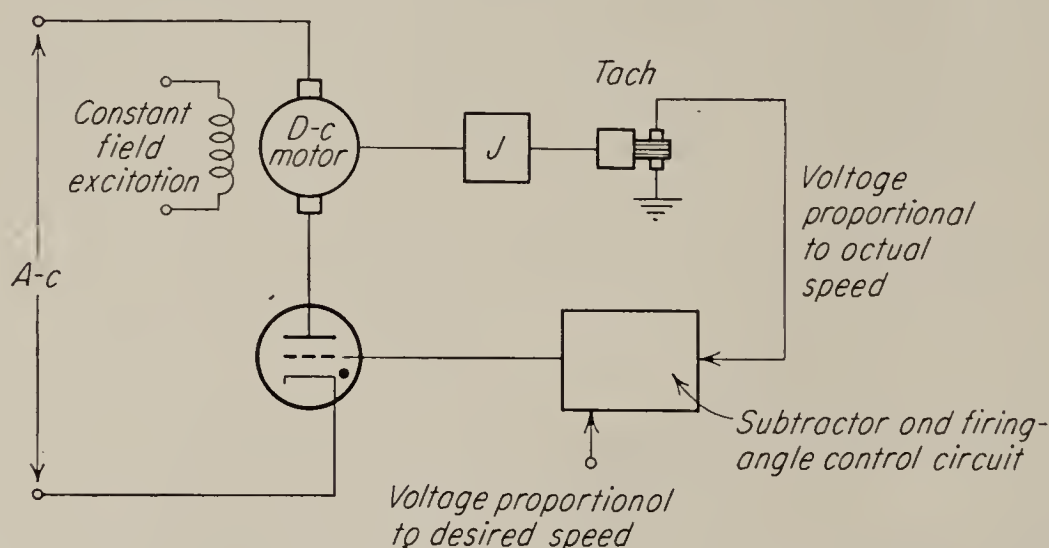


FIG. 3.26. Block diagram of typical thyatron motor-control system.

may slow down too much before the thyatron fires again and may then again overspeed and coast. The action resembles the bouncing of a ball being pushed along a horizontal plane by a block moving at uniform speed.

By differentiation of Eqs. (3.47) and (3.48) it can be shown that the slope of the curves for zero developed torque is zero. This means that the speed regulation near zero torque is very great; i.e., the motor must slow down considerably to support a relatively small torque change. On the other hand, near full torque the curves are all approximately linear. They may, therefore, be used in determining a quasi-linear transfer function, as described in Sec. 4.16. In general, the effect of the thyatron amplifier can be approximated by a series resistance which tends to increase the motor time constant.

3.10. The Effect of Armature Inductance. In the preceding section we have considered the armature inductance of the motor to be negligible. However, since the ratio of inductive reactance at line frequency to the resistance is the significant parameter determining the effect of the inductance, this assumption is not always completely justified. In

general, however, it is true that the inductance, even if not completely negligible, is small.

An accurate analysis of the effect of inductance was shown in Sec. 3.6 to be a rather involved task. However, if the inductance is small, its effect may be taken into account approximately without too much difficulty. We assume, therefore, that the inductance is sufficiently small so that $2\pi fL_a \ll R_a$, where f is the line frequency.

During the period of tube conduction, the armature current may be thought of as consisting of two components, an a-c component due to the a-c supply voltage and a d-c component. If the inductance is small, the a-c component will be approximately sinusoidal in shape, particularly toward the end of the conducting period when the initial switching transient can be assumed to have decayed. Also, if the inductive reactance is small compared to the resistance, the amplitude of the a-c component is approximately E_s/R_a . Thus, the only effect of the inductance on the a-c component of the armature current is to delay it, so that its cycle is behind the a-c supply voltage by the angle $\tan^{-1}(2\pi fL_a/R_a)$. The d-c component of the current cannot be affected by the inductance at all. Hence, we find that the instantaneous armature current during the period of tube conduction is approximately

$$i_a = \frac{E_s}{R_a} \sin \left(2\pi ft - \tan^{-1} \frac{2\pi fL_a}{R_a} \right) - \frac{E_g}{R_a} \quad (3.49)$$

Once the thyatron has been fired, it continues to conduct until the plate current goes through zero. By setting the instantaneous current equal to zero in Eq. (3.49), we find that the extinction point is given by

$$2\pi ft = \phi'_b = \sin^{-1} \frac{E_g}{E_s} + \tan^{-1} \frac{2\pi fL_a}{R_a} \quad (3.50)$$

Thus note that to a first approximation the effect of a small amount of inductance in the armature circuit is to delay the extinction point by the angle $\tan^{-1}(2\pi fL_a/R_a)$. Since the motor-terminal voltage is equal to the applied a-c voltage during tube conduction and is equal to E_g during periods of nonconduction, the waveshape of the terminal voltage has the form shown in Fig. 3.27, which shows operation in mode 2.

As can be seen from the figure, the delay in the extinction point reduces the area under the applied-voltage curve, and it therefore reduces both the average voltage and current. The reduction in area is equal to the

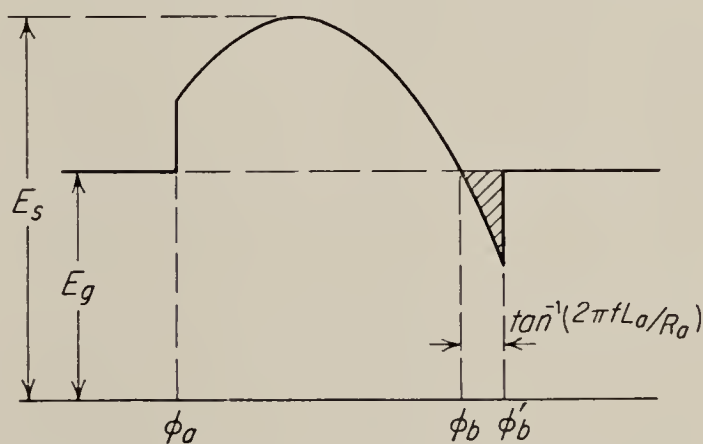


FIG. 3.27. Waveshape of motor-terminal voltage when inductance is present in the armature.

section shown shaded in Fig. 3.27. If $\tan^{-1} (2\pi f L_a / R_a)$ is small, this shaded area is approximately triangular in shape and is therefore approximately equal to

$$\Delta A = \frac{1}{2} \left(\tan^{-1} \frac{2\pi f L_a}{R_a} \right)^2 \frac{d}{d\phi} (E_s \sin \phi) \Big|_{\phi=\phi_b} \quad (3.51)$$

But since $\phi_b = \pi - \sin^{-1} E_g / E_s$,

$$\begin{aligned} \frac{d}{d\phi} (E_s \sin \phi) \Big|_{\phi=\phi_b} &= E_s \cos \left(\pi - \sin^{-1} \frac{E_g}{E_s} \right) \\ &= - \sqrt{E_s^2 - E_g^2} \end{aligned} \quad (3.52)$$

Also for small values of the argument, $\tan^{-1} x \approx x$. Therefore, the shaded area in Fig. 3.27 is approximately equal to

$$\Delta A = - \frac{1}{2} \left(\frac{2\pi f L_a}{R_a} \right)^2 \sqrt{E_s^2 - E_g^2} \quad (3.53)$$

and the reduction in the average current is therefore

$$\Delta I = \frac{\Delta A}{2\pi R} = - \frac{1}{4\pi R_a} \left(\frac{2\pi f L_a}{R_a} \right)^2 \sqrt{E_s^2 - E_g^2} \quad (3.54)$$

The reduction in current will, of course, result in reduced torque. It is clear from Eq. (3.54) that the reduction depends on E_g and therefore on the speed. It is zero when the motor runs at maximum speed and is largest at standstill.

3.11. D-C Motor Driven by Full-wave Circuit. For smoother operation of the motor, the full-wave circuit shown in Fig. 3.28 is often pre-

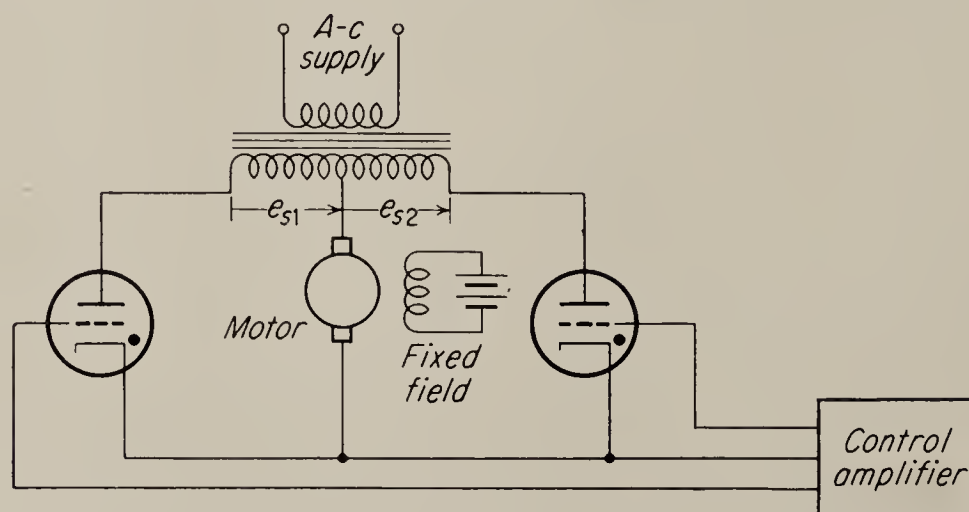


FIG. 3.28. Full-wave thyatron circuit with motor load.

ferred over the half-wave circuit discussed in Sec. 3.10. The operation of the circuit is quite similar to the half-wave circuit as long as the current is not continuous. In fact, all that has been said concerning operating modes, speed-torque curves, effect of inductance, etc., applies to the full-wave circuit, except that, for a given firing angle and speed, the armature

current and therefore the torque are twice as great as in the half-wave circuit. However, as was pointed out in Sec. 3.7, under certain conditions the current in the full-wave circuit may be continuous. If the inductance is negligible, this can happen only for negative speeds and is therefore rather unlikely during normal operation. However, if the armature inductance is sufficiently large, the current may become continuous if the speed is low enough and the firing angle small enough. The extension of the theory to this case is straightforward and is left to the reader.

3.12. Thyatron Circuits for Reversing Control. Two commonly used circuits permitting bidirectional motor-speed control are shown in Fig.

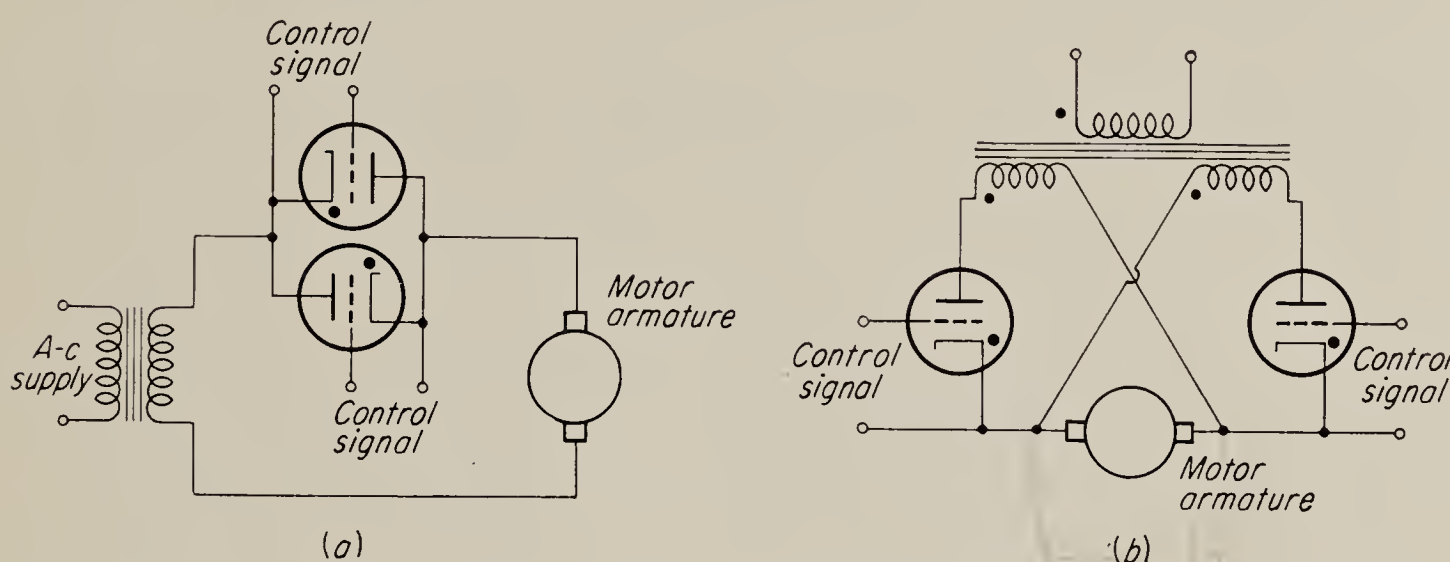


FIG. 3.29. Thyatron circuits for bidirectional motor-speed control. Both are half-wave circuits.

3.29. Both of these circuits are half-wave circuits in which one tube controls the motor when it is running in one direction while the other tube controls it in the opposite direction. The transformer connections in the circuit of Fig. 3.29b are such that the two tubes conduct on alternate half cycles. This is necessary, since there would otherwise be a short circuit through the tubes and transformer windings whenever both tubes conducted simultaneously. This problem does not exist in the simpler circuit of Fig. 3.29a; however, in it also the two tubes can conduct only on alternate half cycles. For smooth transfer of control from one tube to the other near zero motor speed, the grid bias is usually so adjusted that for zero signal input both tubes fire for a short part of their respective half cycle. An increase of signal in the positive direction advances the firing angle of one tube, say, tube 1, and retards the angle of the other tube (tube 2), until eventually only tube 1 conducts. For negative signal the action is reversed.

When only one tube conducts, the analysis of the circuit is exactly the same as that carried through in Sec. 3.9. However, in the interval near zero speed the effect of the two tubes operating together must be con-

sidered. For the sake of simplicity we assume that the armature inductance and the tube drop are negligible.

The waveshape of the armature current is shown in Fig. 3.30 for operation in mode 2. The average armature current is found by integrating over the two pulses of current occurring per cycle. We have, therefore,

$$\begin{aligned}
 I_{av} &= \frac{1}{2\pi R_a} \int_{\phi_{a1}}^{\pi - \sin^{-1} E_g/E_s} (E_s \sin \phi - E_g) d\phi \\
 &\quad + \int_{\phi_{a2}}^{2\pi + \sin^{-1} E_g/E_s} E_s (\sin \phi - E_g) d\phi \\
 &= \frac{1}{2\pi R_a} [E_s (\cos \phi_{a1} + \cos \phi_{a2}) - E_g (3\pi - \phi_{a1} - \phi_{a2})] \quad (3.55)
 \end{aligned}$$

Suppose now that $\phi_{a1} = \phi_{a0} - \Delta\phi_a$ and $\phi_{a2} = \pi + \phi_{a0} + \Delta\phi_a$. This implies that for zero signal the firing angle of tube 1 is ϕ_{a0} and that the

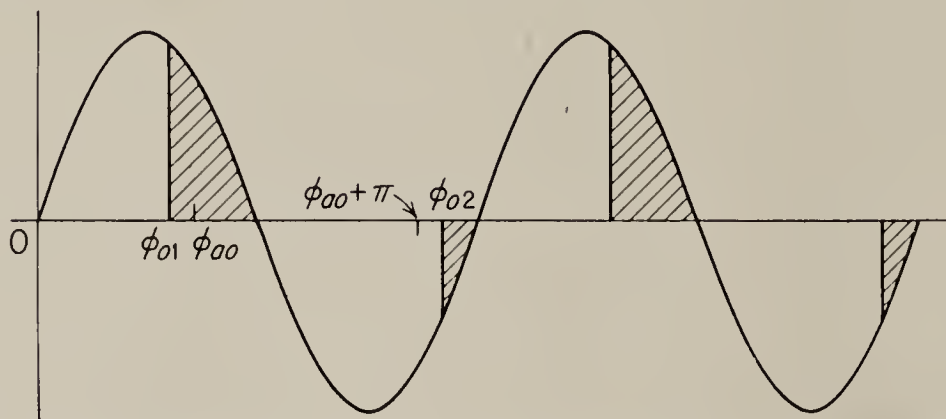


FIG. 3.30. Armature-current waveshape for the circuit of Fig. 3.29.

firing angle of tube 2 comes 180° later. Also the signal $\Delta\phi_a$ advances the firing angle of one tube by the same amount that it retards the firing angle of the other. Substituting into Eq. (3.55), we obtain after some reduction

$$I_{av} = \frac{1}{\pi R_a} [E_s \sin \phi_{a0} \sin \Delta\phi_a - E_g (\pi - \phi_{a0})] \quad (3.56)$$

Since the average developed torque $Q_d = k_t I_{av}$ and since the counter emf $E_g = k_v \Omega$, as in Sec. 3.9, we convert Eq. (3.56) into a relation between speed and torque:

$$Q_d = \frac{k_t}{\pi R_a} [E_s \sin \phi_{a0} \sin \Delta\phi_a - k_v \Omega (\pi - \phi_{a0})] \quad (3.57)$$

This relation holds only when both tubes are conducting and are in mode 2. If it is assumed that the grid-control circuit limits variations of firing angles to the range between 0 and 180° on each tube, then Eq. (3.57) applies only if $\Delta\phi_a$ meets all of the following conditions:

$$\begin{aligned}
 0 &< \phi_{a0} - \Delta\phi_a < \pi \\
 0 &< \phi_{a0} + \Delta\phi_a < \pi \\
 \sin^{-1} \frac{k_v \Omega}{E_s} &< \phi_{a0} - \Delta\phi_a < \pi - \sin^{-1} \frac{k_v \Omega}{E_s} \\
 -\sin^{-1} \frac{k_v \Omega}{E_s} &< \phi_{a0} + \Delta\phi_a < \pi + \sin^{-1} \frac{k_v \Omega}{E_s}
 \end{aligned} \tag{3.58}$$

A composite set of speed-torque curves for the motor can now be plotted by use of Eqs. (3.47), (3.48), and (3.57), keeping in mind the limitations

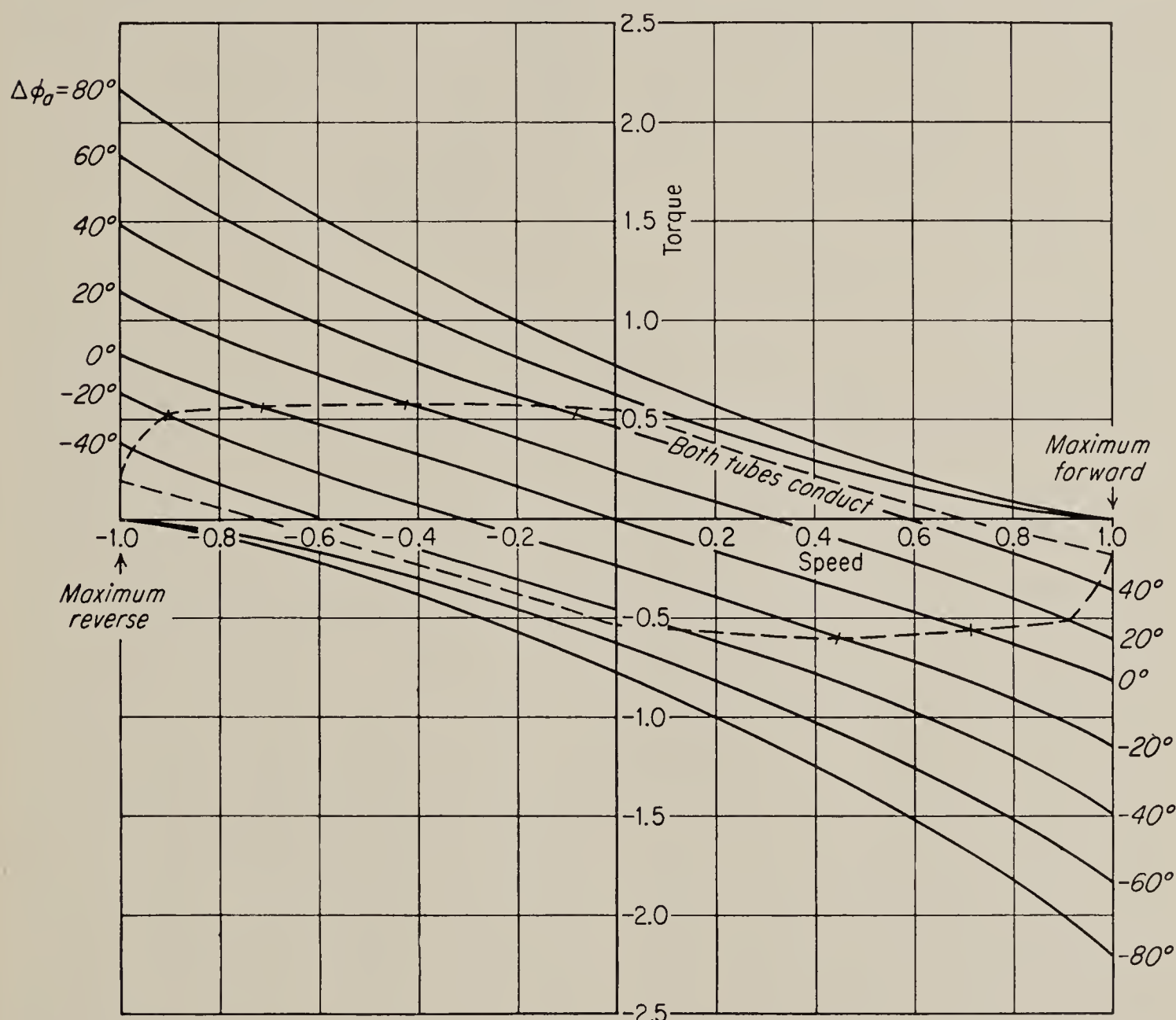


FIG. 3.31. Speed-torque curves for bidirectional thyatron control circuit. Armature inductance and tube drop are neglected; $\phi_{a0} = 135^\circ$.

on speed and firing angles given by inequalities (3.58). Such a set of curves for a quiescent firing angle $\phi_{a0} = 135^\circ$ is given in Fig. 3.31. Of particular interest in this figure is the fact that, in the region of zero speed and small $\Delta\phi_a$, the speed-torque curves are parallel straight lines. The slope of the curves in this region can be obtained by differentiating Eq. (3.57) with respect to Ω ; the result is

$$\frac{dQ_d}{d\Omega} = -\frac{k_t k_v}{R_a} \left(1 - \frac{\phi_{a0}}{\pi} \right) \tag{3.59}$$

Comparison of the speed-torque curves of a d-c motor driven by a linear source (Sec. 4.16) indicates that the thyatron-controlled motor in its linear region is equivalent to a linear d-c motor with a resistance equal to $R_a/(1 - \phi_{a0}/\pi)$. Hence it has a time constant

$$T_M = \frac{JR_a}{k_v k_t (1 - \phi_{a0}/\pi)} \quad (3.60)$$

It is clear that the time constant is reduced by making ϕ_{a0} as small as possible. A small value of ϕ_{a0} also has the advantage of extending the region of linearity of the circuit. However, ϕ_{a0} should not be made less than 90° , since this would reduce the maximum speed available from the motor. A disadvantage of small ϕ_{a0} is that the efficiency is reduced and the motor heating is increased. Thus, if efficiency and heating are important factors, as they would be in large power installations, ϕ_{a0} must be kept as large as possible.

When ϕ_{a0} is near 90° , the circuit may operate in mode 1, as described in Sec. 3.9, and become nonlinear from that cause. Since such small values of ϕ_{a0} are used relatively rarely, the analysis of this type of operation will not be carried through here. It offers no particular difficulty.

The effect of armature inductance may be considered for the bidirectional circuit in the same way as for the unidirectional circuit. If the inductance is small, then the reduction of current in the positive pulse is approximately canceled by the reduction in the negative pulse, particularly at low speeds. Thus the effect tends to be even less for the bidirectional circuit than for the unidirectional circuit.

Full-wave bidirectional circuits are seldom used because of the difficulties of avoiding short-circuit paths. Unless special precautions, e.g., current-limiting resistances, are employed to limit the short-circuit current, a full-wave circuit could never be operated with the firing points of the forward and backward halves of the circuit overlapping. The circuit would therefore be required to have a deadband region in which none of the tubes conduct. This, however, is very undesirable in a servomechanism, in which operation about the null is the rule.

It should be noted that all the foregoing analyses of thyatron circuits have considered only average behavior of the circuits. In control systems designed for a sufficiently wide frequency band to require significant amplification at frequencies near line frequency, the pulsed nature of the thyatron current must be considered. The thyatron amplifier is then best analyzed on the sampled-data basis.¹

3.13. Other Thyatron Applications. Although the control circuits discussed in detail up to this point are of fundamental importance,

¹ Truxal, "Automatic Feedback Control System Synthesis," McGraw-Hill Book Company, Inc., New York, 1955, chap. 9.

there are a few other circuits using thyratrons which deserve brief mention.

The split-field series motor discussed in Sec. 4.19 is readily adapted to thyatron drive. A simple half-wave control circuit for this service is shown in Fig. 3.32. When thyatron 1 conducts for a longer period than

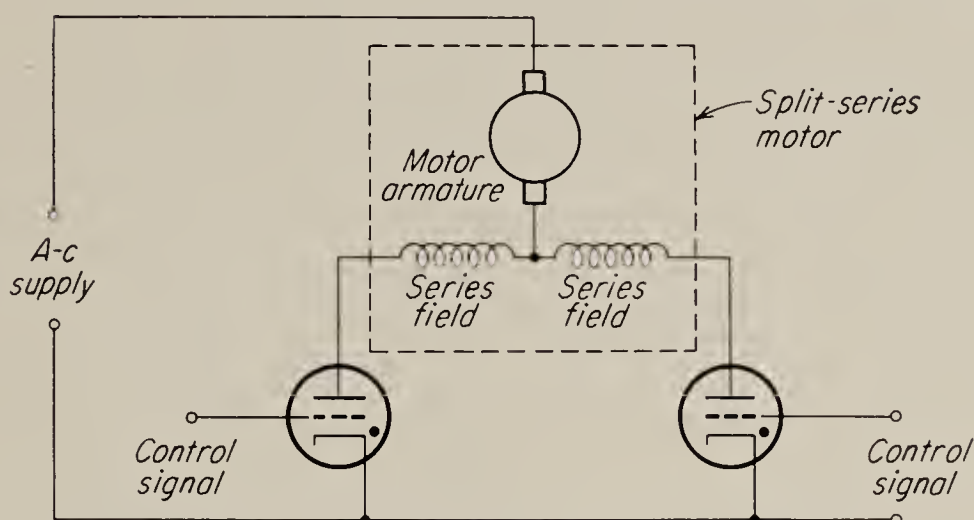


FIG. 3.32. Thyatron control circuit for split-field series motor.

thyatron 2, the motor turns in one direction, whereas if thyatron 2 conducts for the longer period, the motor reverses. At the expense of somewhat greater complexity of the control circuits, the circuit can be converted to full-wave operation.

Thyratrons may be used to control the operation of a-c motors of both the induction type and the commutator type.¹ In Fig. 3.33 is shown a two-phase induction servomotor driven from a single-phase supply. When the forward thyratrons are allowed to conduct, phase 1 is driven directly and phase 2 is driven through the phase-splitting capacitor C . Allowing the reverse thyratrons to conduct reverses the phase excitation and thus the direction of rotation. As the firing angle of the thyatron pair is retarded, the rms voltage applied to the motor is smoothly reduced to essentially zero.

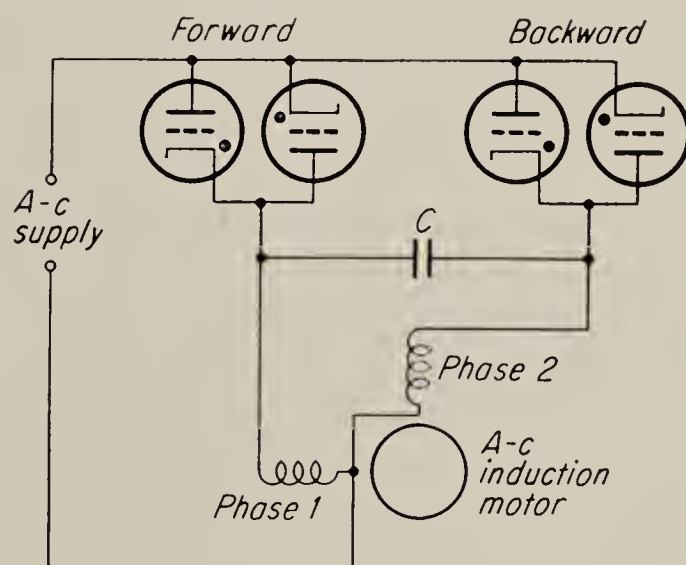


FIG. 3.33. Thyatron control circuit for two-phase induction servomotor.

An example of control of a commutator type of motor by means of thyratrons is given in Fig. 3.34. An a-c repulsion-type motor is shown, in which one or the other set of brushes is short-circuited by a conducting thyatron, depending on the direction of rotation desired.

¹ J. H. Burnett, Driving Servomotors with Grid-controlled Thyratrons, *Elec. Mfg.*, vol. 50, p. 138, October, 1952.

In applications of thyratrons a number of practical points must be kept in mind. Some of these, such as the maximum current and voltage ratings of the various electrodes, are assumed to be familiar to the reader.¹ Other points of this sort, such as electrical noise generated by thyratrons, etc., are discussed in a number of books on industrial electronics and practical control.² An interesting cause of oscillation in thyatron circuits is discussed by Greenwood, Holdam, and MacRae.³

Since grid current is drawn by a conducting tube, the grid bias of the nonconducting tube is changed by the resulting voltage drop in the common grid circuit. As the conducting tube cuts off, the operating points

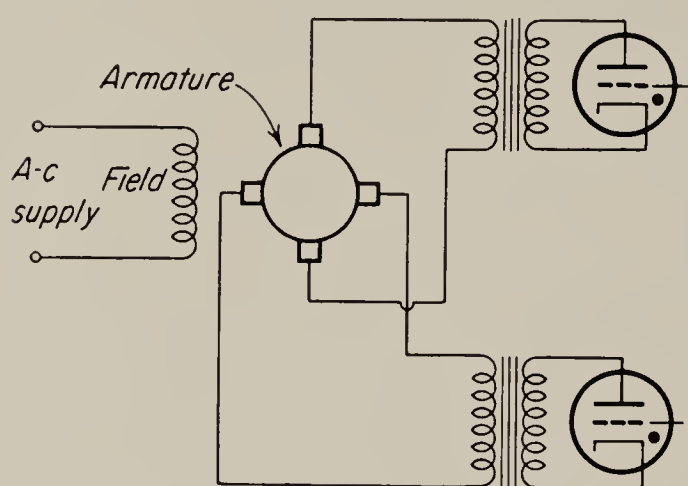


FIG. 3.34. An example of thyatron control of commutator-type motor.

shift, and this may result in a spurious error signal and continuous oscillation of the output device. This might occur in the circuit of Fig. 3.29b, for instance, if the preceding amplifier has a significant output impedance. The remedy is to isolate the control circuits for the two tubes by decreasing the output impedance of the preceding stage or by providing separate tubes to drive the thyatron grids.

3.14. Relay Amplifiers.⁴ The electromagnetic relay is one of the simplest and least expensive power amplifiers available. Its chief disadvantage is that it represents a highly nonlinear circuit element. By use of negative feedback, it is, however, possible to obtain an almost perfectly linear power amplifier still retaining many of the advantages of simple relays.

The basic configuration is given in Fig. 3.35a. The operation may be described by assuming that a step of voltage is applied to the input at time zero. Before this time the output of the relay was zero. The input step begins to charge the capacitor, and at t_1 in Fig. 3.36 the voltage e_r has reached a large enough value to close the relay. At t_1 , then, e_m suddenly becomes E_b ; that is, the contact of the relay closes, and the output is connected to the positive supply. Now the capacitor must charge to $e_i - E_b$,

¹ However, see Dow, "Fundamentals of Engineering Electronics," John Wiley & Sons, Inc., New York, 1945, and Reich, "Theory and Applications of Electron Tubes," McGraw-Hill Book Company, Inc., New York, 1944, and other texts cited earlier.

² Ahrendt, "Servomechanism Practice," McGraw-Hill Book Company, Inc., New York, 1954, p. 114.

³ Greenwood, Holdam, and MacRae, "Electronic Instruments," Radiation Laboratory Series, vol. 21, McGraw-Hill Book Company, Inc., New York, 1948, p. 411.

⁴ J. Gibson and F. Tuteur, The Response of Relay Amplifiers with Feedback, *Trans. AIEE*, vol. 76, part II, pp. 303-307, 1957.

and a downward exponential begins. At t_2 the capacitor voltage has become so small that the relay cannot hold in, and the contact falls open. The capacitor once more charges to a value large enough to close the relay, and the process continues. The relay chatters, and the output is a chain of pulses. This is certainly not a faithful reproduction of the input,

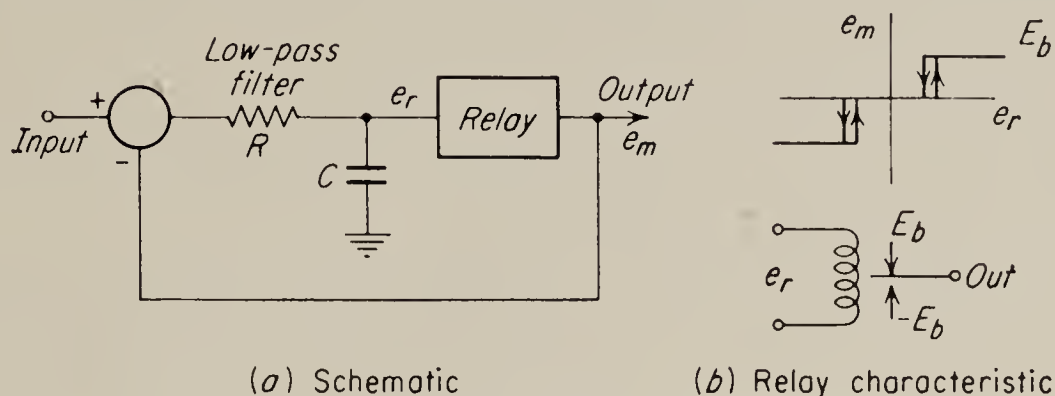


FIG. 3.35. A relay amplifier with feedback.

but if the load acts as a low-pass filter and is essentially sensitive only to the average value of the pulse chain, it is shown below that the amplifier is quite satisfactory and in fact "linear" over most of the operating range. If the input-signal amplitude is large, the pulses are long and the average value of the output is high. The limit on input-signal amplitude is reached at the point at which the relay remains closed for the entire period.

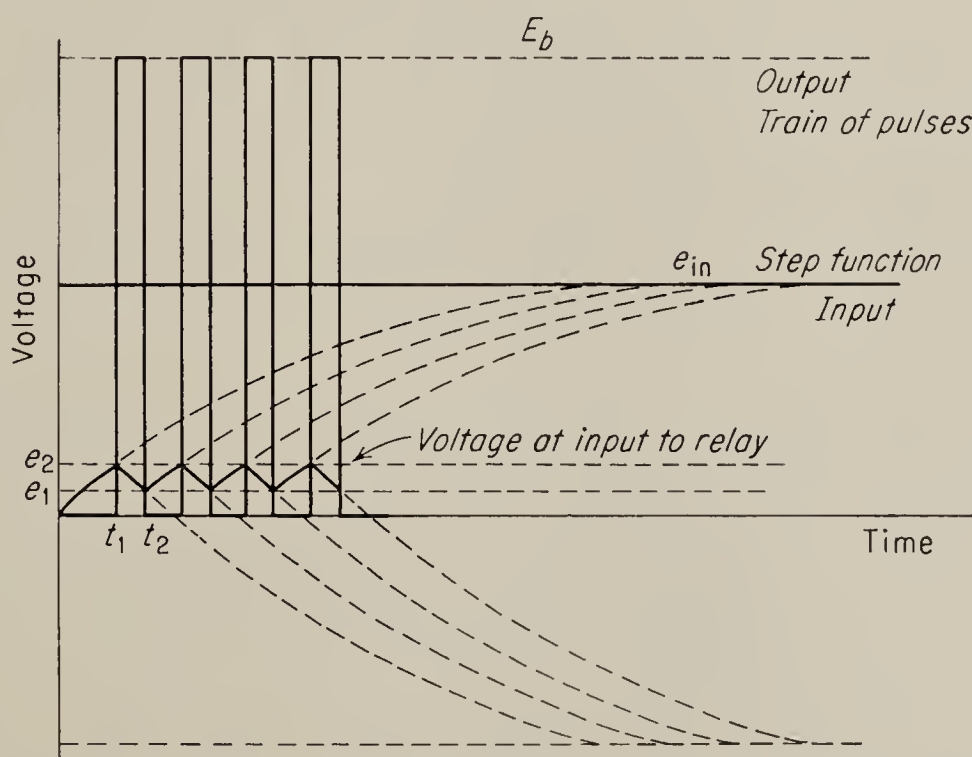


FIG. 3.36. Waveshapes for relay amplifier with step-function input.

The relay used in the amplifier must be a polarized relay, i.e., one that will close to one contact for a positive voltage applied to the coil and to the other contact for a negative voltage applied. The same effect may, however, be obtained by using two unpolarized relays which are in conjunction with diodes. In any physical relay there is a deadband, or

range of coil voltage that is too small to activate the relay. All practical relays also display hysteresis; i.e., a larger value of coil voltage is required to close the relay than that required as a minimum to hold in the relay.

The major advantages of relay amplifiers are simplicity, ruggedness, and economy. The major disadvantage is the fact that the load must supply a good deal of smoothing if the operation is to be acceptable. However, it is shown below that, for a typical motor with an integration and a single time constant, a good replica of the input to the relay amplifier is possible at the shaft output.

3.15. Relay-amplifier Stability. Before discussing the input-output relations of a relay amplifier with feedback, we consider the stability of

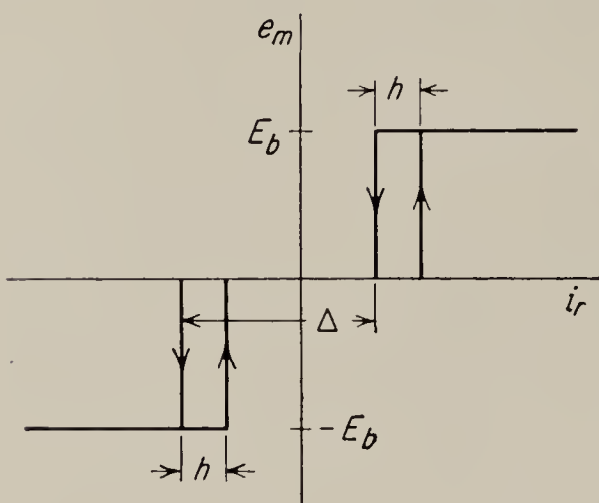


FIG. 3.37. Relay characteristic.

such a device. The stability of feedback systems containing relays is perhaps most easily investigated by use of the describing-function method of Kochenburger.¹ In this method the highly non-linear input-output relation of the relay is replaced by a *quasi-linear describing function*. This describing function is the relation between the magnitude of a sinusoidal input signal and the magnitude and phase angle of the fundamental component of the output. Kochen-

burger has shown that a relay with the input-output characteristics shown in Fig. 3.37 has the describing function

$$H_2 = \frac{E_{m1}}{I_r} = \frac{4 \sin \beta}{\pi I_r} \angle -\alpha \quad (3.61)$$

where E_{m1} is the complex magnitude of the fundamental component of the output voltage, I_r is the input amplitude, and

$$\begin{aligned} \alpha &= \frac{1}{2} \left(\cos^{-1} \frac{\Delta - h}{2I_r} - \cos^{-1} \frac{\Delta + h}{2I_r} \right) \\ \beta &= \frac{1}{2} \left(\cos^{-1} \frac{\Delta - h}{2I_r} + \cos^{-1} \frac{\Delta + h}{2I_r} \right) \end{aligned} \quad (3.62)$$

In Eqs. (3.62) Δ and h are the deadband and hysteresis, respectively, of the relay, as indicated in Fig. 3.37.

The remainder of the feedback loop making up the system is assumed to be linear and has the transfer function H_1 . The system is stable if

¹ R. J. Kochenburger, A Frequency Response Method for Analyzing and Synthesizing Contactor Servomechanisms, *Trans. AIEE*, vol. 69, p. 270, 1950.

the roots of the characteristic equation $1 + H_1 H_2 = 0$ lie in the left half plane for all possible values of H_2 . Stability can be determined by a modification of the usual Nyquist method. In its more commonly used form, this modification takes the form of a complex plot of $-H_2$ and $1/H_1$, as shown in Fig. 3.38 for the case of a relay feedback system in which H_1

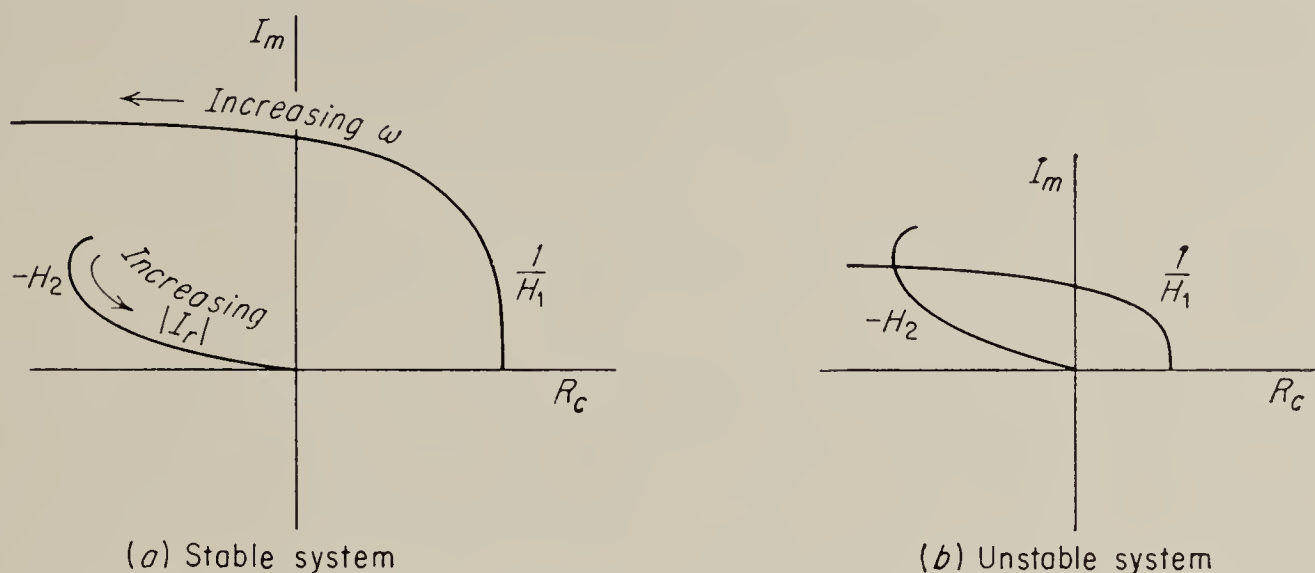


FIG. 3.38. Typical inverse Nyquist diagrams for a feedback system containing a relay.

has at least two lags. Note that H_2 is a function of I_r , while H_1 is a function of frequency. Intersection of the two curves indicates instability.

For the relay amplifier with feedback shown in Fig. 3.39, H_1 is given by

$$H_1 = \frac{kR_1}{(R_1 + R_2) \left[1 + \frac{R_1 R_2 C s}{R_1 + R_2} \right]} = \frac{k'}{1 + T s} \quad (3.63)$$

where k is the amplifier gain. Therefore the Nyquist plot will have the form shown in Fig. 3.40. Since the curve of $1/H_1$ does not intersect that

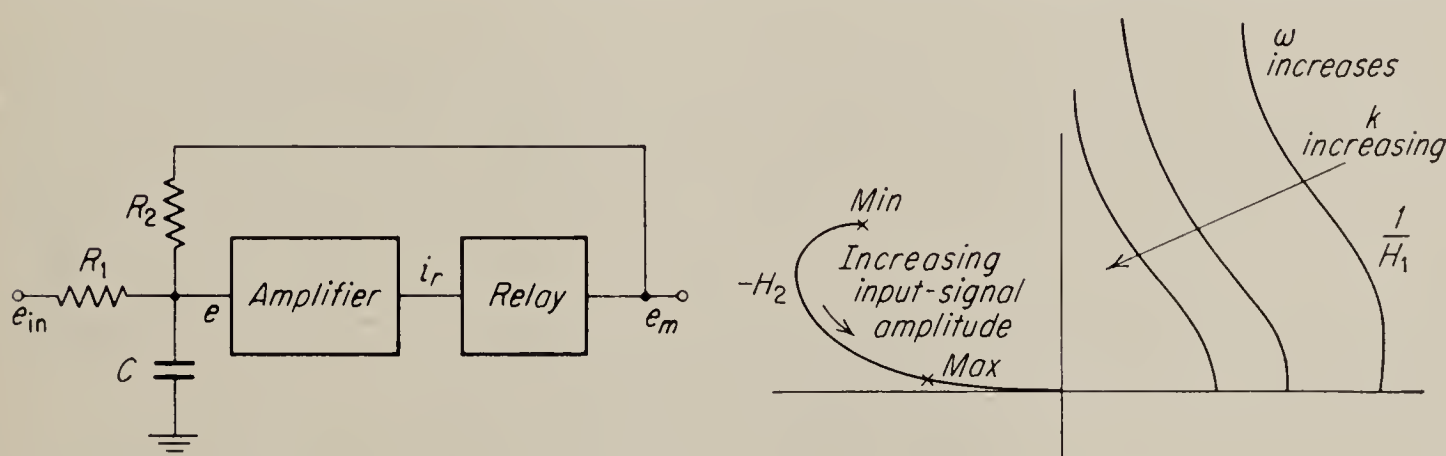


FIG. 3.39. Typical relay-amplifier circuit. FIG. 3.40. Inverse Nyquist diagram for relay amplifier of Fig. 3.39.

for $-H_2$, the implication is that the system is stable no matter how large the gain k is. However, at high gains, previously neglected lags, e.g., the time required for the relay to close after the coil is energized, may cause the frequency locus to bend around as shown in Fig. 3.38, thus possibly causing instability. Similarly, if the filter break frequency is raised to a

value such that other lags in the loop, such as the mechanical and/or electrical lags of the relay itself, become important, the frequency locus bends to the left around the origin, as in the case of increased gain. If it intersects the amplitude locus, instability is indicated.

3.16. Relay-amplifier Static Characteristic. The output of a relay amplifier with feedback is a series of pulses, positive or negative, of amplitude E_b and with length and repetition rate determined by the input.

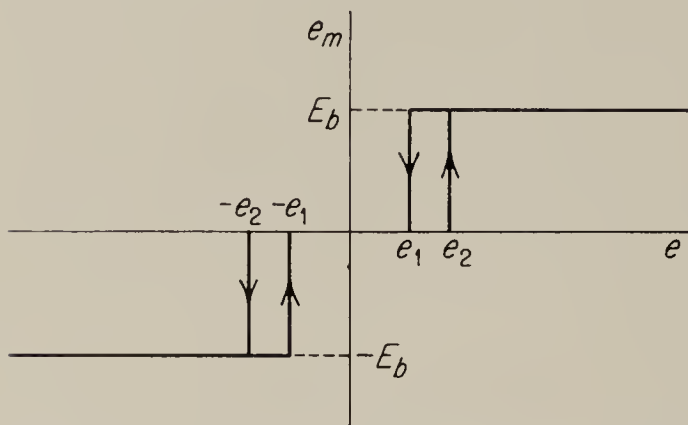


FIG. 3.41. Pull-in and drop-out voltages defined on the relay characteristic.

Usually the relay amplifier is employed in an application where only the average value of this train of pulses is important.

In order to determine the static relation between the input voltage e_{in} and the average output voltage E_m , we shall consider the circuit in Fig. 3.39, with the static characteristic between e and e_m as shown in Fig. 3.41. Let us suppose that, at

zero time, t_0 , $e = 0$. Thus the relay is open, and $e_m = 0$. Let a step of voltage, E_{in} in magnitude, be applied, as shown in Fig. 3.36. Then

$$e = E_{in}(1 - e^{-t/T_1}) \quad (3.64)$$

where $T_1 = R_1C$. When e reaches e_2 , the relay closes. Substituting e_2 for e in Eq. (3.64), we may solve for t_1 , the time of closing:

$$E_{in} - e_2 = e^{-t_1/T_1} \quad (3.65)$$

$$\text{or} \quad t_1 = -T_1 \ln (E_{in} - e_2) \quad (3.66)$$

After the relay closes,

$$e = E_{in} \frac{R_2}{R_1 + R_2} [1 - e^{-(t-t_1)/T_2}] - E_b \frac{R_1}{R_1 + R_2} [1 - e^{-(t-t_1)/T_2}] + e_2 e^{-(t-t_1)/T_2} \quad (3.67)$$

where $T_2 = [R_1R_2/(R_1 + R_2)]C$. When e reaches e_1 , the relay opens. The time of opening, t_2 , may be found by setting $e = e_1$ and solving for $t = t_2$:

$$e^{-(t_2-t_1)/T_2} = \frac{e_1 - E_{in} \frac{R_2}{R_1 + R_2} + E_b \frac{R_1}{R_1 + R_2}}{e_2 - E_{in} \frac{R_2}{R_1 + R_2} + E_b \frac{R_1}{R_1 + R_2}} \quad (3.68)$$

$$\text{Thus} \quad t_2 - t_1 = -T_2 \ln \frac{e_1 - E_{in} \frac{R_2}{R_1 + R_2} + E_b \frac{R_1}{R_1 + R_2}}{e_2 - E_{in} \frac{R_2}{R_1 + R_2} + E_b \frac{R_1}{R_1 + R_2}} \quad (3.69)$$

With the relay open,

$$e = E_{\text{in}}[1 - e^{-(t-t_2)/T_1}] + e_1 e^{-(t-t_2)/T_1} \quad (3.70)$$

It will be noted that Eq. (3.70) is the same as Eq. (3.64) plus the term due to the initial voltage on the capacitor. All the segments of the voltage curve will be exactly the same as those described by Eqs. (3.67) and (3.70). Equation (3.64) is the special case when no charge is on the capacitor. The relay closes again at t_3 , when $e = e_2$; this time is given by

$$t_3 - t_2 = -T_1 \ln \frac{e_2 - E_{\text{in}}}{e_1 - E_{\text{in}}} \quad (3.71)$$

During the period of time $t_3 - t_2$ the output voltage is $-E_b$, and during the period $t_2 - t_1$ the output is zero. These two lengths of time represent the cycle of the output-voltage wave. The average output is therefore

$$E_m = \frac{-E_b(t_2 - t_1)}{t_3 - t_1} \quad (3.72)$$

$$\text{or } E_m = \frac{E_b \ln \frac{e_1 - E_{\text{in}} \frac{R_2}{R_1 + R_2} + E_b \frac{R_1}{R_1 + R_2}}{e_2 - E_{\text{in}} \frac{R_2}{R_1 + R_2} + E_b \frac{R_1}{R_1 + R_2}}}{\left(\frac{T_1}{T_2} \ln \frac{e_2 - E_{\text{in}}}{e_1 - E_{\text{in}}} + \ln \frac{e_1 - E_{\text{in}} \frac{R_2}{R_1 + R_2} + E_b \frac{R_1}{R_1 + R_2}}{e_2 - E_{\text{in}} \frac{R_2}{R_1 + R_2} + E_b \frac{R_1}{R_1 + R_2}} \right)} \quad (3.73)$$

By a relatively minor change in the circuit, R_2 can be grounded whenever the relay is open. This will make $T_1 = T_2$ and will contribute to the linearity of the input-output relation. It is also usual to make $R_1 = R_2$. In this case Eq. (3.73) reduces to

$$E_m = \frac{E_b \ln \frac{1 + 2e_1/(E_b - E_{\text{in}})}{1 + 2e_2/(E_b - E_{\text{in}})}}{\ln \frac{1 - E_b/(E_{\text{in}} - 2e_1)}{1 - E_b/(E_{\text{in}} - 2e_2)}} \quad (3.74)$$

In certain applications an electronic difference amplifier is used rather than the resistor-adding network. The low-pass filter may be incorporated as a part of the interstage coupling network of the electronic amplifier. When the difference amplifier rather than the resistor-adding network is employed, the factors representing the voltage-divider action do

not appear, and Eq. (3.73) can be simplified to yield

$$E_m = \frac{E_b \ln \frac{1 + e_1/(E_b - E_{in})}{1 + e_2/(E_b - E_{in})}}{\ln \frac{1 - E_b/(E_{in} - e_1)}{1 - E_b/(E_{in} - e_2)}} \quad (3.75)$$

By examination of this relation we can see that it gives the required negative

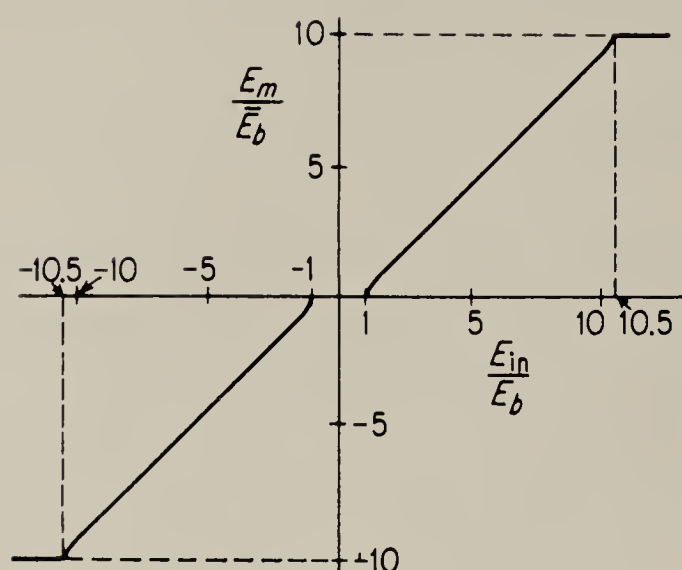


FIG. 3.42. Static characteristic of the relay amplifier, normalized to the relay-closing voltage E_b . The contact voltage was ten times the closing voltage in this normalization.

real result only for the region $e_2 < E_{in} < E_b + e_1$. For $E_{in} < e_2$ the relay does not close; the output is zero. For $E_{in} > E_b + e_1$, the relay remains closed and the output is E_b . It may be shown that, if e_1 and e_2 are small with respect to the battery voltage, the slope of E_m versus E_{in} is essentially constant over the operating range. The exact relation between input-voltage magnitude and the average of the output voltage can be determined by the use of Eq. (3.75). In Fig. 3.42 is shown E_m versus E_{in} normalized to E_b .

3.17. The Frequency Response of a Relay Amplifier.

The limitation on the frequency response for a sinusoidal input of a relay amplifier is determined to a first approximation by the time constant T of the filter or the *chatter rate* of the output pulses. Actually, however, this chatter rate or *output-pulse repetition rate* depends on the input-signal magnitude, and in the limit, where E_{in} is either greater than E_b or zero, the relay remains always closed or always open. Since the relay amplifier is a nonlinear device, the entire concept of frequency response is contrived, and a measure for one type of input signal will not hold for another. For a step-function input, the relay amplifier is particularly fast and essentially establishes the required average value of e_m in one opening and closing.

The response of a relay amplifier to sinusoidal signals depends on the amplitude and the frequency of the input signal. For signals whose peak amplitude at the input of the relay is less than the value required to close the relay, there is no output. For instance, if the closing voltage of the relay is 10 volts, a signal whose peak amplitude is less than 10 volts will not close the relay. Since the low-pass filter attenuates the amplitude of the input signal, the peak amplitude must be computed at the input to the relay. At the break frequency of the filter, for instance, the filter

reduces the input amplitude by a factor of 0.707; thus the minimum input signal amplitude must be 14.14 volts peak. Essentially a relay amplifier will operate properly for normal-amplitude signals up to the break frequency of the low-pass filter. In Fig. 3.43 are shown the amplitude reduction and phase shift for a typical relay amplifier whose relay-closing voltage is 10 per cent of the battery voltage applied to the relay terminals. The frequency response shown in Fig. 3.43 was obtained by smoothing the pulse-train output in order to recover the input-signal frequency component. When the input-signal frequency is equal to or greater than the

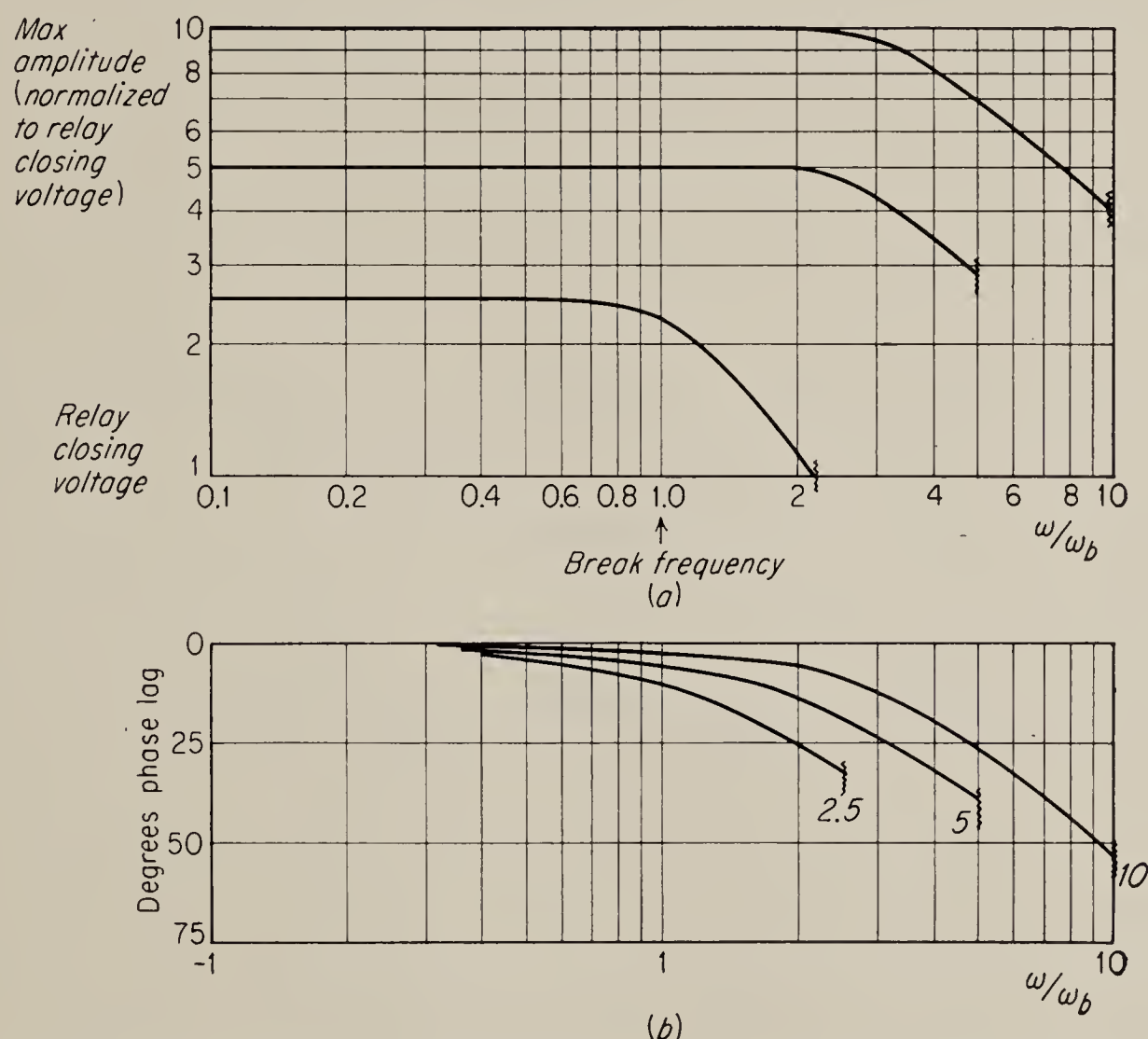


FIG. 3.43. Frequency response of relay amplifier normalized to relay-closing voltage.

break frequency of the filter, especially at low amplitudes, the number of output pulses per cycle becomes small, and even the integrated output becomes rather steplike. In Fig. 3.44 the various waveshapes for a particular signal are shown. Fig. 3.44a shows the sine-wave input to the amplifier; its normalized peak amplitude is 5.0; its frequency is the same as the break frequency of the amplifier low-pass filter. In Fig. 3.44b is shown the waveform at the output of the low-pass filter, which illustrates the deadband and hysteresis of the relay. In Fig. 3.44c is given the output of the relay, and Fig. 3.44d shows the integrated output. In a typical power-amplifier application the relay output drives a d-c motor, which would represent an integration plus a low-pass filter. Thus it could be

expected that the waveshape in d would be smoothed by the filter before appearing at the output shaft.

It is difficult to establish an equivalent loop gain constant for a relay amplifier with feedback without making several rather arbitrary assumptions. However, from Fig. 3.43 it can be seen that such an equivalent gain depends on the input-signal amplitude. At maximum-amplitude signal the break frequency of the closed-loop response of the amplifier occurs at about $4/T$, and for minimum-amplitude signals the break approaches $1/T$.

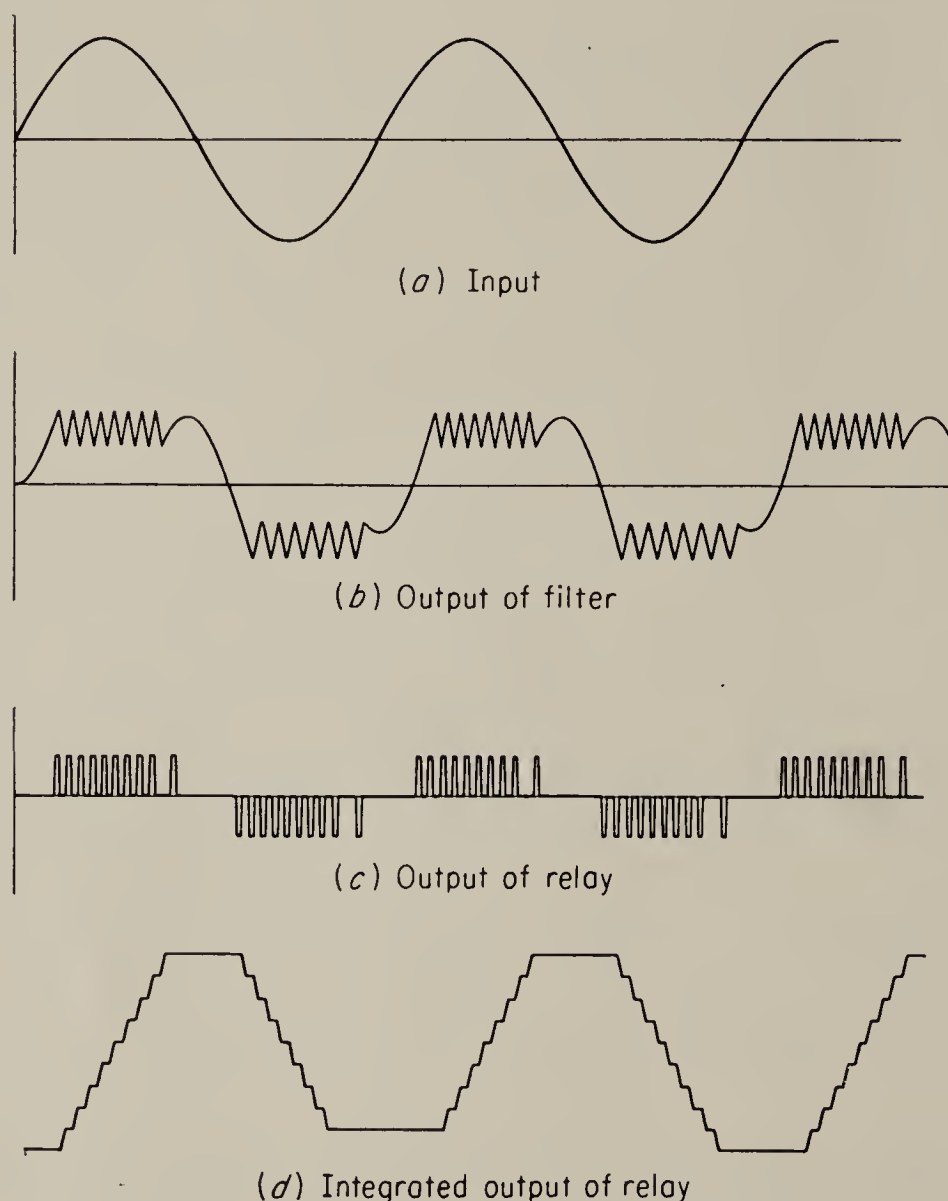


FIG. 3.44. Relay-amplifier waveshapes.

The amount of phase shift contributed by the filter to the over-all phase shift shown in Fig. 3.43 is inversely related to the gain of the amplifier, as a result of the negative feedback. In addition to decreasing the closed-loop phase shift, an increase in amplifier gain will act to reduce the effective deadband, thus increasing the usefulness of the amplifier at low signal amplitudes. As the gain is increased, the loop becomes less stable from the increased effect of previously neglected lags in the amplifier and relay itself, until finally the amplifier goes into oscillation, as shown in the preceding section. That is, its output for no input is a continuous string of alternate positive and negative pulses. As a signal is

placed on the input terminals, the length of the positive pulses grows longer and that of the negative pulses shorter, or vice versa; thus the averaged output follows the input. This type of operation is more sensitive to low-amplitude signals than is stable operation and will have minimum phase shift from input to output. However the continuous vibration of the power element under quiescent conditions may be undesirable.

3.18. Magnetic Amplifiers. Introduction. Magnetic amplifiers have become more and more widely used in recent years for both high- and low-power applications. The simplest form of magnetic amplifier requires a d-c input signal and delivers an a-c output. This type may be used when the load is one that requires alternating current, such as the a-c servomotor (see Chap. 7). In conjunction with rectifiers, magnetic amplifiers also can deliver a d-c output signal and may then be used to drive the field of a d-c generator, the solenoid of a hydraulic valve, etc. In power-handling capacity, magnetic amplifiers range from very small sizes capable of delivering a few milliwatts to very large sizes designed for outputs of many kilowatts.

The primary advantages of magnetic amplifiers over electron-tube amplifiers are ruggedness and the absence of filaments that require warm-up time and constitute an important source of power loss. Magnetic-amplifier circuits can be designed for a wide range of input and output impedances, and since the control and load circuits are not conductively coupled, they provide somewhat more flexibility than vacuum-tube circuits. On the other hand, magnetic amplifiers have a limited frequency response and produce an output that is a modulated a-c carrier. They have a much lower input impedance than vacuum tubes and provide only a finite power gain, whereas vacuum tubes may, in theory at least, give an infinite power gain. A practical disadvantage is that magnetic amplifiers must in general be designed specially for each application, since both the supply voltage and the load impedance critically affect the operation. Hence magnetic amplifiers are usually considerably more expensive than vacuum tubes, particularly when they are used in applications where mass production economies are not applicable.

A simple magnetic-amplifier circuit is shown in Fig. 3.45a. This circuit is the series-connected type and consists of two iron cores wound with two windings each. The iron cores are made of special alloys such as *Delta-max* or *Supermalloy* which saturate sharply and with a relatively low level of magnetizing force H . A typical hysteresis loop of a standard core material is shown in Fig. 3.45b. One of the windings on each of the cores is referred to as the load or gate winding. In the series amplifier the two gate windings are connected in series with the load to the a-c source. The other two windings, the control windings, are also connected in series to the control source; however they are connected in opposition, so that the

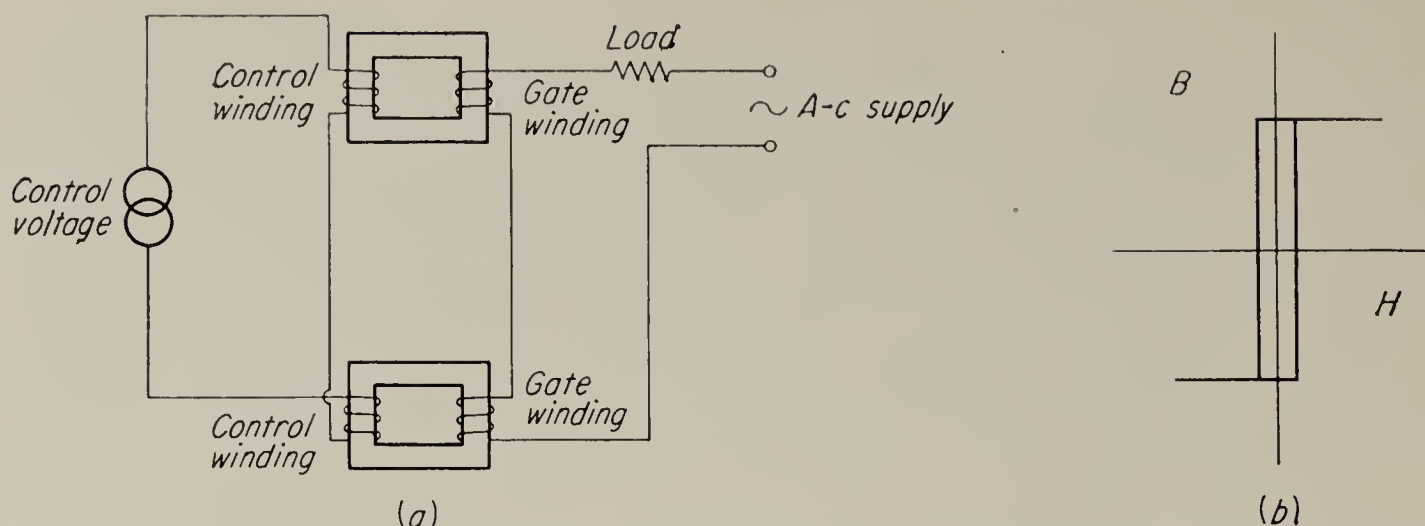


FIG. 3.45. Series-connected magnetic amplifier, and hysteresis loop of core material.

fundamental component of a-c voltage induced in the control windings is canceled and does not affect the control source.

A number of other core and coil arrangements are also used in practice, depending on the application.

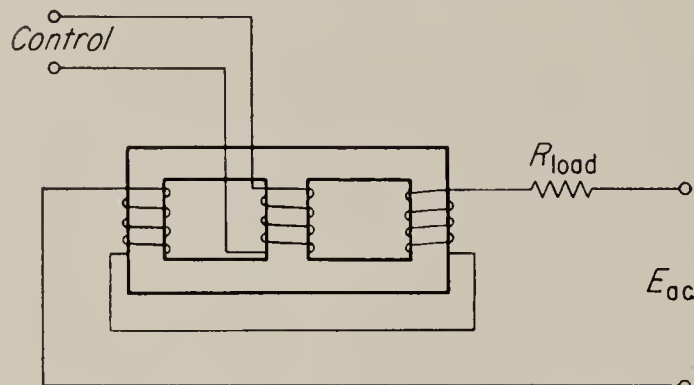


FIG. 3.46. Magnetic amplifier using three-legged core.

Thus the two gate windings may be connected in parallel, as shown in Fig. 3.53. Quite often a three-legged core is used, as shown in Fig. 3.46, and frequently the cores are wound with more than one control winding. The basic operating principle is, however, essentially the same for all of these types, and it will therefore be illustrated by using

the series-connected circuit as an example.

3.19. Operation of the Series-connected Magnetic Amplifier. In order to simplify the explanation of the operation of a series-connected circuit as much as possible, the following assumptions are usually made:

1. The core has either zero or infinite inductance, depending on the state of the flux in the core.
2. The resistance of the control circuit including the control source is very small, although not always negligible.
3. The gate windings have zero resistance.
4. The load is resistive.

Assumption 1 amounts to a simplification of the hysteresis loop of the core material to three straight lines at right angles to each other, as shown in Fig.

3.47, and corresponds fairly closely to the hysteresis curves found in cores made of high-quality iron. Assumption 4 is made in this necessarily elementary treatment because the operation of magnetic amplifiers

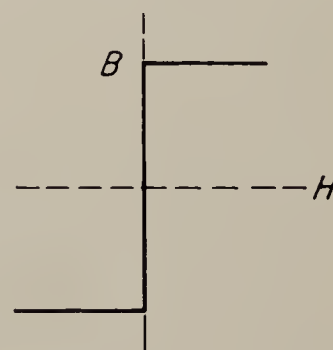


FIG. 3.47. Idealized hysteresis loop.

with reactive loads is fairly complicated and beyond the scope of this text.¹ Since the first three assumptions are obviously never exactly met in practice, it may be expected that the following analysis will not be exact; however it is found that the results obtained correspond reasonably well to those obtained in actual operation.

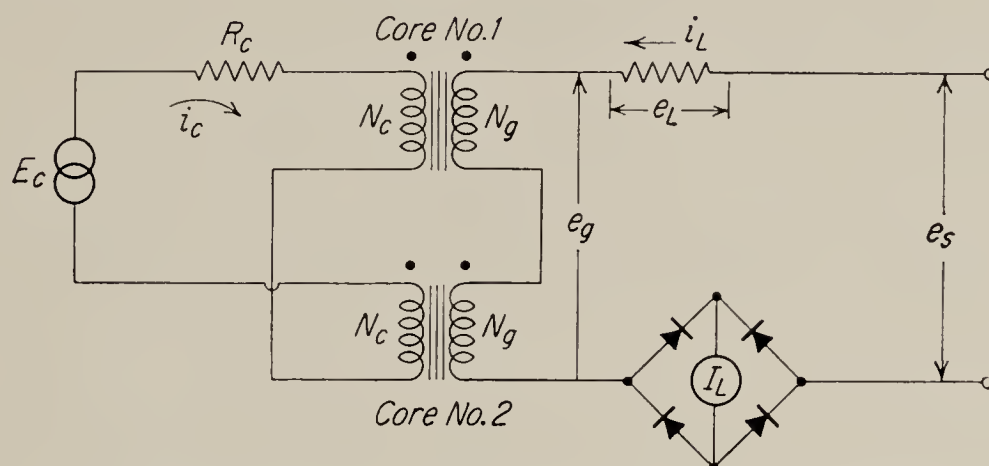


FIG. 3.48. Schematic diagram of a magnetic amplifier.

We consider first the operation of a magnetic amplifier with zero control current. The alternating supply voltage applied to the two gate windings sets up an alternating flux in the cores, which in turn generates a counter voltage of self-induction, as given by Faraday's law:

$$\frac{1}{2} e_g = N_g \frac{d\phi}{dt} \quad (3.76)$$

where e_g is the total voltage appearing across both gate windings and N_g is the number of turns in each gate winding. By assumption 1 no current is required in the winding to establish the flux. Hence there is no current in the load and no voltage drop across the load, so that the reactor voltage e_g is at all times exactly equal and opposite to the applied voltage e_s (see Fig. 3.48 for the notation used). A reactor is said to be *normally excited* when the maximum flux set up in the core by the applied a-c supply voltage is exactly equal to the saturation value ϕ_s . A normally excited reactor never saturates if the control current is zero, and the load current remains zero during the entire a-c voltage cycle. The flux and voltage cycles obtained under these conditions are shown in Fig. 3.49.

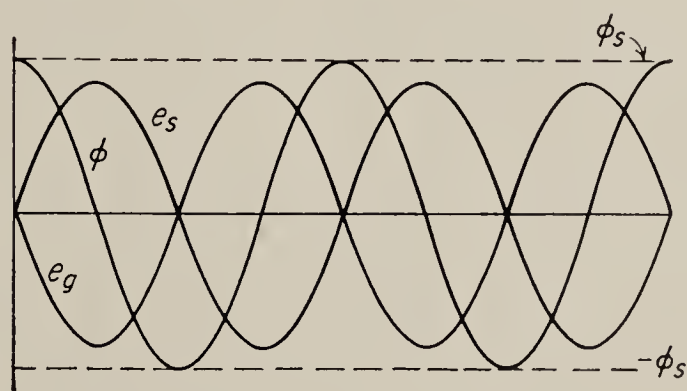


FIG. 3.49. Flux and voltage in a normally excited reactor with zero control current.

When control current is passed through the control windings, an addi-

¹ See, however, Storm, "Magnetic Amplifiers," John Wiley & Sons, Inc., New York, 1955.

tional magnetomotive force is added to each core and causes the cores to become saturated for some part of the a-c cycle. Suppose that, at a certain point in the cycle, reactor 1 of Fig. 3.48 saturates. When this happens, the flux can no longer change, or

$$\frac{d\phi}{dt} = 0 \quad (3.77)$$

Hence the reactor can no longer generate any counter emf and, by assumptions 1 and 3, now represents a short circuit, not only on its gate winding, but also on its control winding. As a result the control winding of reactor 2 is now connected only across the resistance of the control circuit, which, by assumption 2, is negligibly small. The entire control circuit seen by reactor 2 therefore represents a short circuit, and even though this reactor is not saturated, the voltage across both of its windings must also be zero. The flux in the core must therefore be constant. Hence all the supply voltage appears across the load. During the next half cycle the process repeats, with the roles of the two reactors reversed. In Fig. 3.50 are shown the resulting waveforms of voltage, flux, and current. In drawing these curves, we assume that the circuit is operating in the steady state and that the reactors are normally excited. At time $t = 0$, reactor 2 is just coming out of saturation, and the flux in reactor 1 has the value ϕ_a . This value is chosen to meet the requirement that the circuit is in steady-state conditions, i.e., that $\phi_1 = \phi_a$ for $\omega t = 2\pi, 4\pi, \dots$

As time increases from zero, the flux in both reactors increases, with $d\phi/dt$ exactly equal to the value required to make the reactor counter voltage equal and opposite to the applied voltage. This state of affairs continues until $\omega t = \alpha$, at which point reactor 1 saturates. As indicated in the preceding paragraph, the flux in both reactors now remains constant, and the voltage drop across the reactors is zero. The supply voltage therefore appears across the load, and load current flows. As soon as the supply voltage passes through zero, $d\phi/dt$ can again be supplied by both reactors. The reactors therefore are again able to generate counter emf, and the load voltage and current go to zero, until at $\omega t = \pi + \alpha$, reactor 2 saturates again and the cycle repeats.

The waveshape of the load-current pulses is the same as that of the load-voltage pulses, since the load is assumed to be resistive. During the first pulse, i.e., for $\alpha < \omega t < \pi$, reactor 1 is saturated, and its load and control circuits are therefore completely decoupled. The fact that the load current flowing in its gate winding produces mmf has no effect on its control winding. In reactor 2, however, the flux is constant at a value that is less than the saturation value. According to the ideal form of the hysteresis loop assumed for the core material (Fig. 3.47), there can there-

fore be no net magnetomotive force in this reactor. Hence the mmf of the gate winding is opposed by an equal and opposite mmf produced by the control winding, or

$$N_c i_c = N_g i_L \quad (3.78)$$

During the next half cycle the load current flows in the opposite direction, but since reactor 1 is connected in the opposite direction to the control

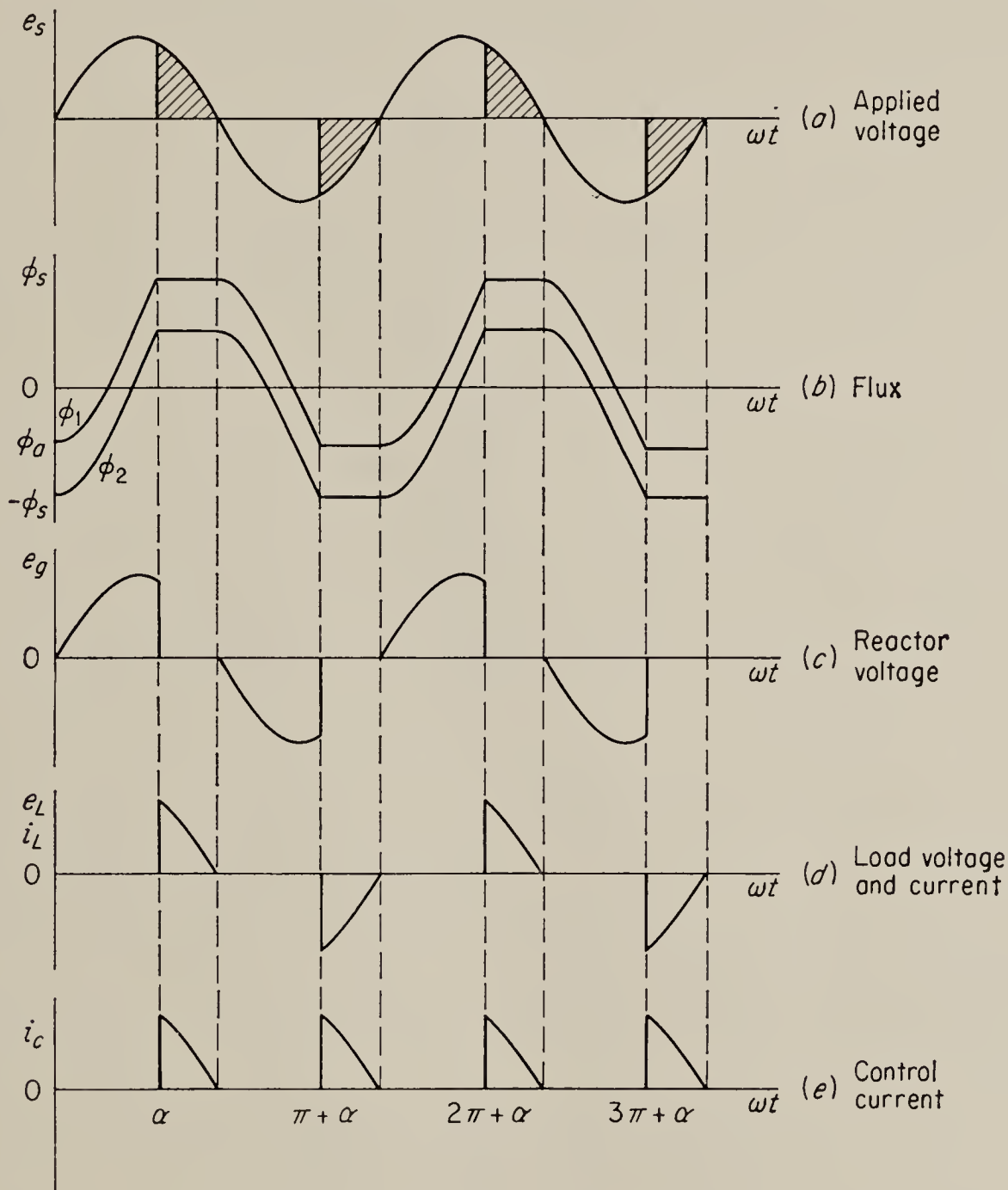


FIG. 3.50. Waveforms in series magnetic amplifier.

circuit, the control-current pulse is in the same direction as the previous pulse. Hence the control current consists of unidirectional pulses, as shown in Fig. 3.50e.

The control current has an average component I_c . Since the wave-shapes of the load- and control-current pulses are the same,

$$N_c I_c = N_g I_L \quad (3.79)$$

where I_L is an average over one half cycle. In Fig. 3.48 is shown one

method of connecting a d-c ammeter and a bridge rectifier to read I_L . Since I_c is a d-c current, which cannot be induced in the control circuit by transformer action, it must actually be the current supplied by the control voltage E_c ; i.e.,

$$I_c = \frac{E_c}{R_c} \quad (3.80)$$

Thus we find that the current gain of the series magnetic reactor is given by

$$K_I = \frac{I_L}{I_c} = \frac{N_c}{N_g} \quad (3.81)$$

where I_c and I_L are average quantities. It is interesting to note that neither the supply voltage nor the various resistances of the circuit enter into this result, and the implication is, therefore, that a magnetic amplifier acts as a controllable constant-current device. This is found to be closely true in actual practice for load currents less than the value requiring continuous conduction of the reactor,

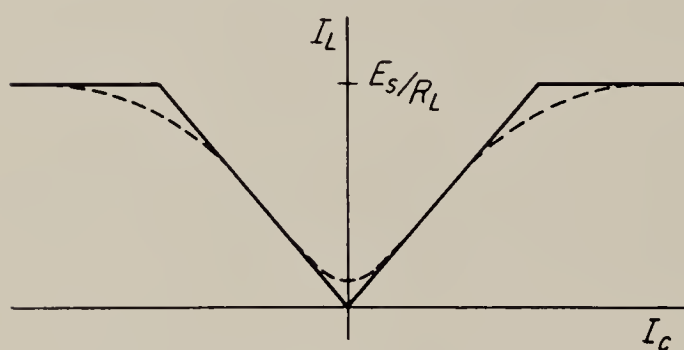


FIG. 3.51. Load current versus control current.

i.e., for $I_L < E_s/R_L$. This is shown in Fig. 3.51, where the solid line shows the theoretical relation and the dotted curve shows a typical measured response. Note that the chief difference is the rounding off at the top and the small current existing for zero control current.

These differences are due to the fact that in actual practice the reactor requires some magnetizing current to establish the flux in the core.

The voltage gain of the series-connected amplifier may be derived from Eq. (3.81):

$$K_E = \frac{E_L}{E_c} = \frac{R_L I_L}{R_c I_c} = \frac{N_c R_L}{N_g R_c} \quad (3.82)$$

Here again, all quantities are average values. The d-c power gain is

$$K_p = \frac{E_L I_L}{E_c I_c} = \frac{N_c^2 R_L}{N_g^2 R_c} \quad (3.83)$$

Since E_L and I_L are average quantities, the power output given by this expression is that obtained after rectifying and filtering the current output and then passing it through the load. If the power content of the a-c current pulses coming from the magnetic amplifier is of interest (as it

might be in a heating or lighting application), the power gain would be

$$K_{p_{ac}} = \frac{N_c^2 R_L}{N_g^2 R_c} F^2 \quad (3.84)$$

where F is the form factor of the current wave shown in Fig. 3.50*d*.

In the preceding discussion, use has been made of the approximation that R_c , the control-circuit resistance, is negligibly small. Since in practical circuits there is always some resistance in the control circuit, the question arises of how large R_c may become, while still being considered negligibly small. As has already been demonstrated, only one reactor saturates at any one time, and the other one is unsaturated. The unsaturated reactor acts exactly like a transformer. The small control-circuit resistance is therefore reflected to the gate winding by means of the standard transformer relation:

$$R'_c = \frac{N_g^2}{N_c^2} R_c \quad (3.85)$$

As long as this reflected resistance is very much less than the load resistance R_L , the voltage drop across the reactors will be negligibly small compared to the load voltage during periods of conduction. A ratio of R_L to R'_c of 10 or more is usually considered adequate to ensure that the assumption of negligible R_c does not result in excessive errors. If R'_c is greater than R_L , the reactors operate in a different mode, which is not considered here.¹

3.20. Time Constant of the Series-connected Amplifier. The relations derived in the previous section were obtained on the basis of steady-state operation. When the control voltage is varied, however, the dynamic performance of the amplifier must also be considered. This performance is affected chiefly by the time constant of the control winding. There is a slight residual time lag in the load winding which is not considered here.

In order to find the time constant of the control winding, we must determine its effective inductance L_c . We may write

$$\begin{aligned} E_c &= R_c i_c + N_c \frac{d\phi_1}{dt} - N_c \frac{d\phi_2}{dt} \\ &= R_c i_c + N_c \frac{d}{dt} (\phi_1 - \phi_2) \end{aligned} \quad (3.86)$$

This may be rewritten as

$$E_c = R_c i_c + N_c \frac{d(\phi_1 - \phi_2)}{di_c} \frac{di_c}{dt} \quad (3.87)$$

¹ *Ibid.*, and Geyger, "Magnetic Amplifier Circuits," McGraw-Hill Book Company, Inc., New York, 1954.

We note by reference to Fig. 3.50b that, for constant control current, $\phi_1 - \phi_2 = \phi_d$ is constant. Hence we may define

$$L_c \triangleq N_c \frac{d\phi_d}{di_c} \quad (3.88)$$

the effective inductance of the control circuit. The problem now is to determine ϕ_d as a function of i_c and the amplifier constants. We proceed by showing first that the average load voltage is directly proportional to ϕ_d .

From the previous discussion,

$$E_L = E_s - E_g \quad (3.89)$$

where E_L , E_s , and E_g are the load, supply, and gate voltages, all averaged over a half cycle of the supply frequency. The instantaneous gate voltage is given by

$$\frac{1}{2} e_g = N_g \frac{d\phi}{dt} \quad (3.76)$$

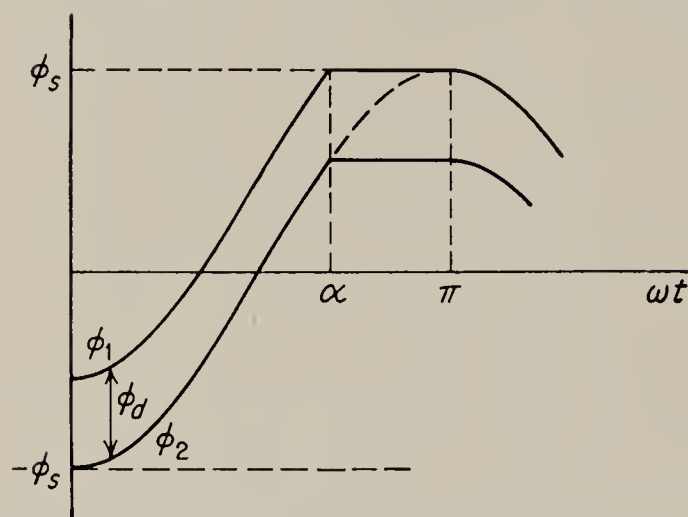


FIG. 3.52. Flux in series-connected reactors.

where ϕ is the instantaneous flux in one of the cores. Suppose the flux in core 2 is as shown in Fig. 3.52. If the core is normally excited, then by inspection of Fig. 3.52 we have

$$\begin{aligned} \phi_2 &= -\phi_s \cos \omega t & 0 < \omega t < \alpha \\ \phi_2 &= -\phi_s \cos \alpha & \alpha < \omega t < \pi \end{aligned} \quad (3.90)$$

Hence

$$\begin{aligned} e_g &= 2N_g \frac{d}{dt} (-\phi_s \cos \omega t) \\ &= 2N_g \omega \phi_s \sin \omega t & 0 < \omega t < \alpha \\ e_g &= 0 & \alpha < \omega t < \pi \end{aligned} \quad (3.91)$$

The average of e_g is found by integration over a half cycle:

$$\begin{aligned} E_g &= \frac{1}{\pi} \int_0^\alpha 2N_g \omega \phi_s \sin \omega t d(\omega t) \\ &= 2N_g \phi_s \frac{\omega}{\pi} (1 - \cos \alpha) = 4N_g \phi_s f (1 - \cos \alpha) \end{aligned} \quad (3.92)$$

where $\omega = 2\pi f$. The average supply voltage E_s is equal to E_g for the case where the reactor does not fire, i.e., where $\alpha = \pi$. Hence

$$E_s = 8N_g \phi_s f \quad (3.93)$$

From Eqs. (3.89), (3.92), and (3.93) we have now

$$\begin{aligned} E_L &= 8N_g \phi_s f - 4N_g \phi_s f (1 - \cos \alpha) \\ &= 4N_g f \phi_s (1 + \cos \alpha) \end{aligned} \quad (3.94)$$

But

$$\phi_s (1 + \cos \alpha) = \phi_s - (-\phi_s \cos \alpha) = \phi_d \quad (3.95)$$

from Fig. 3.52, so that, finally, the desired result is

$$E_L = 4N_g f \phi_d \quad (3.96)$$

All that is required now to find L_c is to relate ϕ_d to i_c . By use of Eqs. (3.85) and (3.86), we have

$$\frac{E_L}{I_c} = \frac{I_L R_L}{I_c} = \frac{N_c R_L}{N_g} \quad (3.97)$$

Hence, from Eqs. (3.86) and (3.96)

$$L_c = \frac{N_c \phi_d}{I_c} = \frac{N_c E_L}{4N_g f I_c} = \frac{N_c^2 R_L}{4N_g^2 f} \quad (3.98)$$

Thus the time constant is given by

$$T_c = \frac{L_c}{R_c} = \frac{N_c^2 R_L}{4f N_g^2 R_c} \quad (3.99)$$

Note that the time constant is inversely proportional to the supply frequency. This indicates that, in order to obtain high response speeds, magnetic amplifiers should be operated from high-frequency supplies.

A commonly used figure of merit applied to power-amplifying equipment is the product of power gain and bandwidth. This figure of merit reduces to a very simple result for magnetic amplifiers if the bandwidth is defined as the reciprocal of the time constant. Using the relation for d-c power gain developed in the previous section [Eq. (3.83)], we find that

$$K_p \frac{1}{T_c} = 4f \quad (3.100)$$

Here again the advantage of high-frequency operation is apparent. It is also clear that, for a given supply frequency, an increase in gain must be paid for by a decrease in bandwidth. This is a result found quite generally in amplifying equipment.

3.21. Reactors Connected in Parallel. The parallel connection is used to supply relatively low voltage, high-current loads, and its operation differs in some respects from that of the series-connected amplifier discussed above. The circuit is shown in Fig. 3.53. In analyzing this circuit, the same assumptions that were made in connection with the series amplifier are employed, except that the assumption of negligible control-circuit resistance is not required. The cores are assumed to be normally excited.

In the parallel connection only the saturated reactor conducts load current. This is true because this reactor represents a short circuit if the gate-winding resistance is assumed to be zero, while the other reactor represents the finite resistance $R'_c = (N_g^2/N_c^2)R_c$. Hence the load-cur-

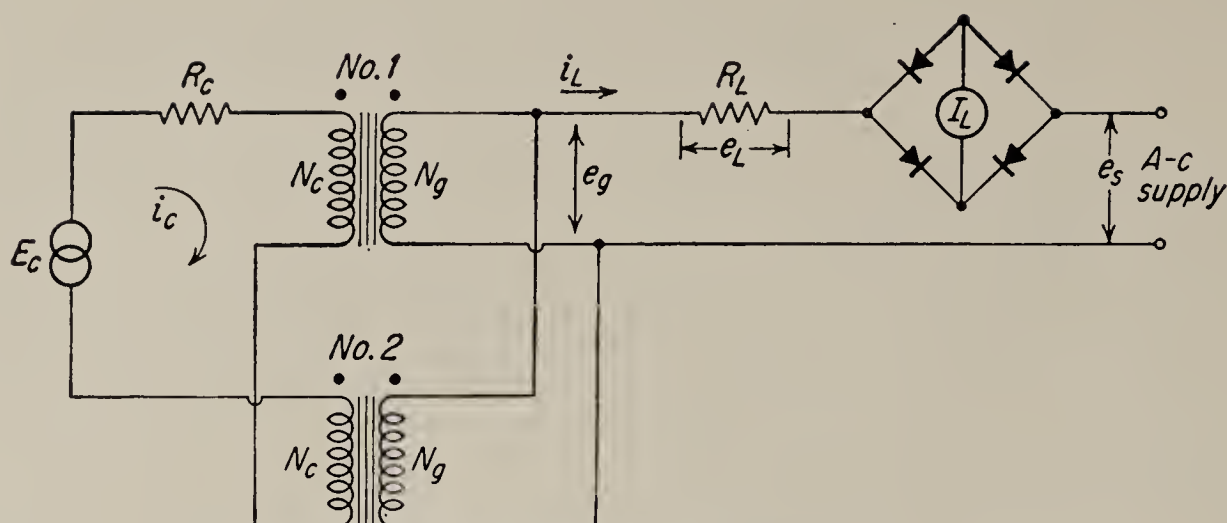


FIG. 3.53. The parallel-connected magnetic amplifier.

rent pulses do not induce pulses in the control circuit, as in the series-connected amplifier, and the control current consists of a steady d-c current. However, by assumption 1 of Sec. 3.19, the net mmf on any core

must be zero during periods when it is not saturated. Hence during these periods there must be a current in the gate winding such that

$$N_g i_g = N_c i_c \quad (3.101)$$

When both reactors are unsaturated, this current flows in opposite directions through the two gate windings. The load current is therefore zero during this time, and only i_g circulates between the two gate windings.

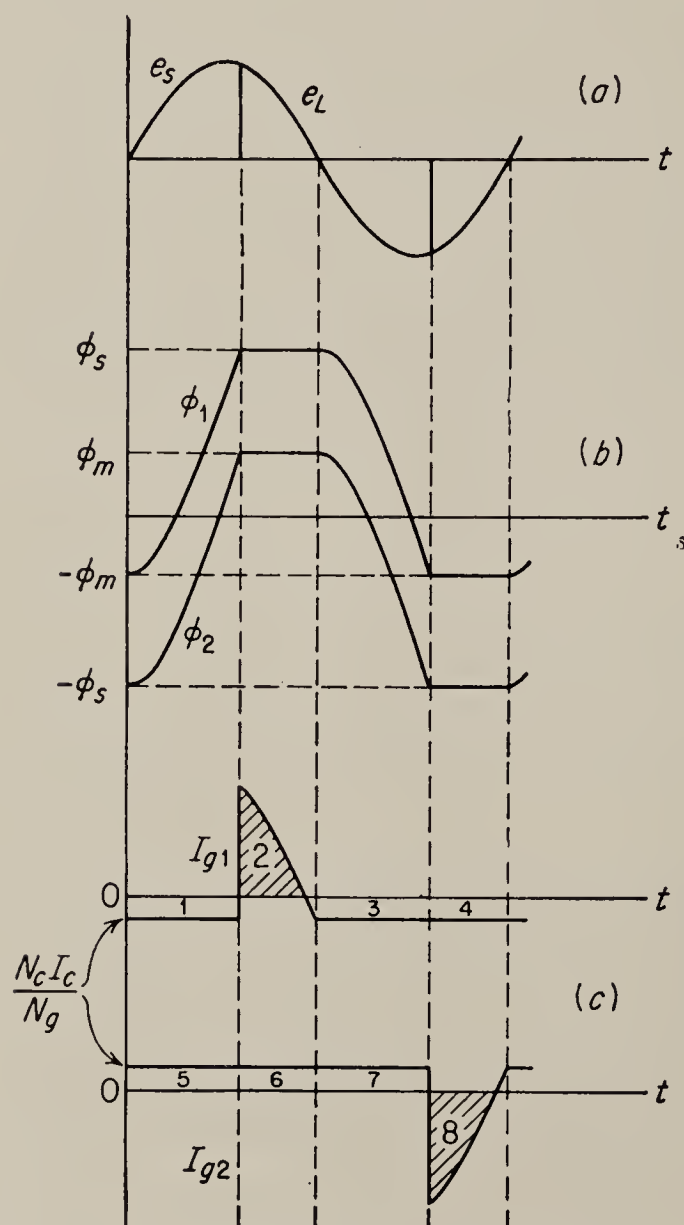
The waveshape of the two gate-winding currents is shown in Fig. 3.54c. The shaded portion represents the load current. The average or d-c value of gate-winding currents must be zero, since there are no rectifiers in the circuit. Hence the average value of a single load-current pulse must be equal to the value of gate-winding current during the period when the reactor is not saturated. For each reactor, therefore,

$$I_L = \frac{N_c I_c}{N_g} \quad (3.102)$$

FIG. 3.54. Waveshapes pertaining to parallel-connected magnetic amplifier.

where I_L is the average value of the load current over half a cycle.

Since each reactor furnishes only half the load-current pulses, the total



load current is given by

$$I_L = \frac{2N_c I_c}{N_g} \quad (3.103)$$

Thus the current gain is given by

$$K_I = \frac{I_L}{I_c} = \frac{2N_c}{N_g} \quad (3.104)$$

The voltage gain is

$$K_e = \frac{E_L}{E_c} = \frac{I_L R_L}{I_c R_c} = \frac{2N_c R_L}{N_g R_c} \quad (3.105)$$

The average or d-c power gain is

$$K_p = \frac{4N_c^2 R_L}{N_g^2 R_c} \quad (3.106)$$

and the rms or a-c power gain is

$$K_{pac} = \frac{4N_c^2 R_L}{N_g^2 R_c} F^2 \quad (3.107)$$

where F is the form factor of the load-current pulses.

The time constant of the parallel-connected amplifier can be obtained by methods similar to those used in connection with the series amplifier.

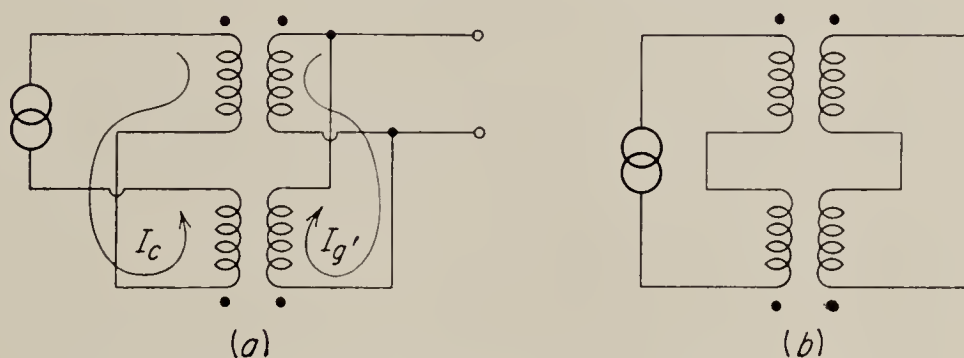


FIG. 3.55. Circulating current in parallel amplifier.

An important difference is that in the parallel circuit the two gate windings permit the flow of a circulating current which increases the total time constant. The path of the circulating current is shown in Fig. 3.55a and may be seen to be equivalent to a short circuit coupled to the control winding. This equivalence is made somewhat clearer in Fig. 3.55b, in which the circuit of Fig. 3.55a has been redrawn with the lower reactors turned upside down and the connections to the load removed.

By means of straightforward circuit analysis¹ it can be shown that, if the coefficient of coupling between the two circuits is unity, the effective time constant is given by

$$T = \frac{L_c}{R_c} + \frac{L_g}{R_g} \quad (3.108)$$

¹ See Sec. 4.3 and Eq. (4.18), where this same problem is discussed in a different connection.

where L_c and L_g are the effective inductances of the control and gate windings, respectively, while R_c and R_g are the corresponding resistances.

The inductance of the control circuit may again be defined as in Eq. (3.88)

$$L_c \triangleq N_c \frac{d\phi_d}{di_c} \quad (3.88)$$

since ϕ_d is a constant for constant i_c as before. By reasoning that is entirely analogous to that used for the series amplifier, we can show that

$$\begin{aligned} E_L &= 2N_g f \phi_d \\ \text{or} \quad \phi_d &= \frac{E_L}{2N_g f} \end{aligned} \quad (3.109)$$

Hence Eq. (3.88) becomes

$$L_c = N_c \frac{d(E_L/2N_g f)}{di_c} = \frac{N_c}{2N_g f} \frac{dE_L}{d(E_c/R_c)} = \frac{N_c R_c}{2N_g f} \frac{dE_L}{dE_c} \quad (3.110)$$

The voltage gain is given by Eq. (3.105) and may be substituted. Hence

$$L_c = \frac{N_c R_c}{2N_g f} \frac{2N_c R_L}{N_g R_c} = \frac{N_c^2 R_L}{N_g^2 f} \quad (3.111)$$

In a like manner L_g may be defined as

$$L_g \triangleq N_g \frac{d\phi_d}{di'_g} = N_g \frac{d(E_L/2N_g f)}{d[(N_c/N_g)i_c]} = \frac{N_g R_c}{2fN_c} \frac{dE_L}{dE_c} \quad (3.112)$$

and, again employing the voltage gain,

$$L_g = \frac{R_L}{f} \quad (3.113)$$

Thus from Eq. (3.108)

$$T = \frac{N_c^2 R_L}{fN_g^2 R_c} + \frac{R_L}{fR_g} \quad (3.114)$$

If the windings are designed according to the usual transformer practice, the resistances of the windings are proportional to the square of the number of turns. If it is then also assumed that the resistance of the control circuit is due only to the resistance of the control winding, we may write

$$R_g = R_c \frac{N_g^2}{N_c^2} \quad (3.115)$$

and the time constant becomes

$$T = \frac{2N_c^2 R_L}{N_g^2 R_c f} \quad (3.116)$$

In order to compare the performance of the parallel- and series-connected

amplifiers, it is again instructive to determine the ratio of d-c power gain to time constant. From Eqs. (3.106) and (3.116) this ratio is found to be

$$\frac{K_p}{T} = \frac{4N_c^2 R_L / N_g^2 R_c}{2N_c^2 R_L / N_g^2 R_c f} = 2f \quad (3.117)$$

Thus it appears that the parallel-connected amplifier has a somewhat poorer performance than the series circuit. It is clear from the preceding analysis that this is due entirely to the short-circuit path existing in the gate-winding circuit.

3.22. Magnetic Amplifiers with Feedback. The performance of magnetic amplifiers can be improved greatly by the use of positive feedback. The feedback can be applied in a number of ways. By way of illustration we consider the series-connected amplifier with feedback applied through a special feedback winding. In a typical circuit such as the one shown in Fig. 3.56, the alternating load current is rectified and applied to feedback windings N_f on each of the cores. The feedback windings are connected in opposition just as the control windings are, and the effect of the current flowing in these windings is the same as that of the current flowing in the control windings. The equal ampere-turn relationship shown to exist in the simple reactor must hold here as well, and we have, therefore,

$$I_L N_g = I_c N_c + I_g N_f \quad (3.118)$$

or

$$I_L = I_c \frac{N_c}{N_g - N_f} \quad (3.119)$$

If we define a feedback ratio B as

$$B \triangleq \frac{N_f}{N_g} \quad (3.120)$$

then

$$I_L = I_c \frac{N_c}{N_g} \frac{1}{1 - B} \quad (3.121)$$

We may also write that

$$E_L = I_L R_L = I_c \frac{N_c R_L}{N_g} \frac{1}{1 - B} = \frac{E_c}{R_c} \frac{N_c}{N_g} R_L \frac{1}{1 - B} \quad (3.122)$$

The current gain is available from Eq. (3.121) and the voltage gain from

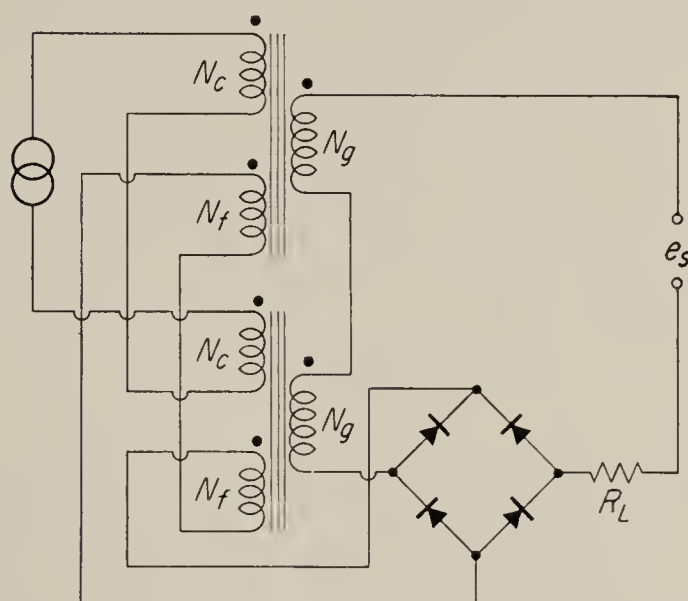


FIG. 3.56. Magnetic amplifier with feedback.

Eq. (3.122). The average d-c power gain is therefore

$$K_p = \frac{E_L}{E_c} \frac{I_L}{I_c} = \frac{R_L N_c}{R_c N_g} \frac{1}{1-B} \frac{N_c}{N_g} \frac{1}{1-B} = \frac{R_L N_c^2}{R_c N_g^2} \frac{1}{(1-B)^2} \quad (3.123)$$

and for the rms power gain we multiply by the square of the form factor, as before.

The time constant of the circuit depends on whether the load current is increasing or decreasing. Inspection of the circuit shows that the rectifier bridge forces the current in the feedback windings to be equal to the load current only when the load current is constant or increasing. When the load current decreases, a larger current may persist in the feedback windings and flow through the rectifier. Thus, for decreasing current the feedback winding constitutes a short-circuited winding which has the same effect on the time constant as the short-circuited gate windings in the parallel amplifier. For rising transients the time constant is therefore L_c/R_c , while for falling transients it is $L_c/R_c + L_f/R_f$.

The inductance L_c may be found as before:

$$L_c \triangleq N_c \frac{d\phi_d}{di_c} \quad (3.88)$$

The relation between ϕ_d and E_L [Eq. (3.96)] is unchanged by the feedback winding; hence, by the same reasoning as that employed in Sec. 3.20,

$$L_c = \frac{N_c^2 R_L}{4N_g^2 f(1-B)} \quad (3.124)$$

The time constant for rising transients is therefore

$$T_c = \frac{L_c}{R_c} = \frac{N_c^2 R_L}{4fN_g^2 R_c(1-B)} \quad (3.125)$$

Since inductance varies as the square of the number of turns, we have

$$L_f = \frac{N_f^2}{N_c^2} L_c \quad (3.126)$$

so that for falling transients the time constant is given by

$$T = \frac{L_c}{R_c} + \frac{N_f^2}{N_c^2} \frac{L_c}{R_f} = \frac{N_c^2 R_L}{4fN_g^2(1-B)} \left(\frac{1}{R_c} + \frac{N_f^2}{N_c^2 R_f} \right) \quad (3.127)$$

The performance of the circuit may again be judged by the figure of merit defined above. From Eqs. (3.123) and (3.125) we have that the ratio of power gain to time constant for rising transients is given by

$$\frac{K_p}{T} = \frac{4f}{(1-B)} \quad (3.128)$$

Thus it is clear that, although the gain and time constant are increased by positive feedback, there is a net advantage in the use of feedback.

The expressions obtained indicate that the gains and time constant become infinitely large as $B \rightarrow 1$. This is not found to be true in practice, primarily because the accuracy of the assumptions made in the derivations begins to diminish. In particular, the assumption of the square hysteresis loop (assumption 1 of Sec. 3.19) becomes less and less tenable, and the actual shape of the hysteresis loop must be considered as the feedback factor approaches unity.

It should also be noted that a tacit assumption made in the above analysis was that the rectifiers are perfect, i.e., that they do not permit any reverse current to flow. In practice, some reverse current always does flow and has the effect of slightly reducing the feedback factor B .

3.23. The Self-saturated Amplifier. The most commonly used method of providing feedback in magnetic amplifiers is the method of *self-saturation*. A number of circuits employing this principle are shown in Fig.

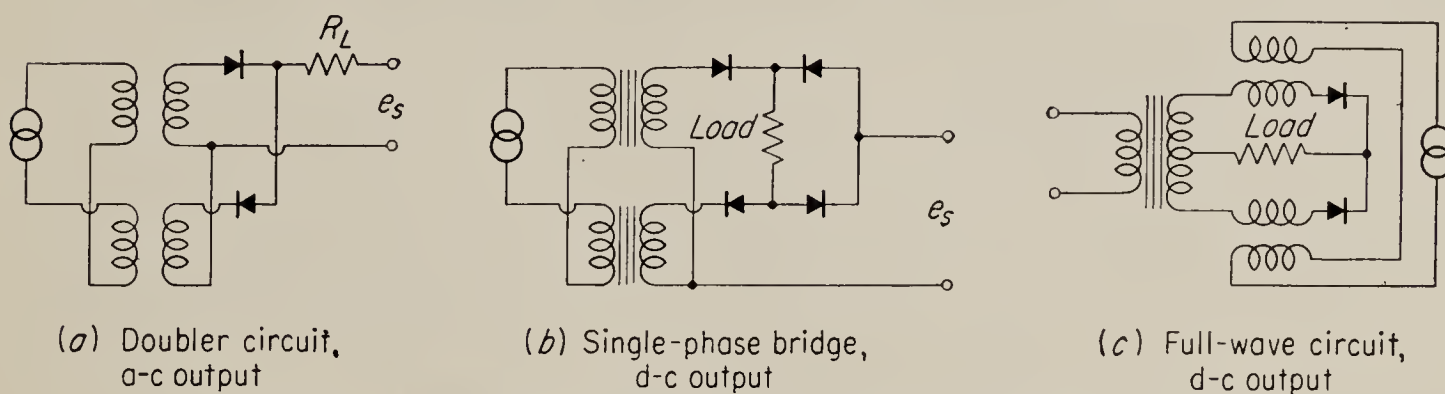


FIG. 3.57. Typical self-saturated magnetic-amplifier circuits.

3.57. In all of these circuits, rectifiers in series with the gate windings produce a d-c component in the gate-winding current which is equal to the average (over one half cycle) of the load current. It can be shown¹ that, if the rectifiers are ideal, i.e., if they do not permit any reverse current flow, the self-saturated amplifier is equivalent to a feedback amplifier in which the feedback factor B is equal to unity. Hence, as indicated above, the simplified analysis of the previous paragraphs no longer yields reasonable results, and a different approach must be used. We present here only an empirical point of view, which takes the experimental control characteristic of the amplifier as its starting point and deduces all further results from it.

A typical self-saturated-amplifier control characteristic is shown in Fig. 3.58a. Note that the load current is almost maximum when the control current is zero and that a "bias" value of negative control current is therefore required to produce minimum output. This is due to the fact that the rectifiers in the circuit cause the amplifier to saturate itself when

¹ W. J. Dornhoefer and V. H. Krummenacher, Applying Magnetic Amplifiers, *Elec. Mfg.*, March, 1951, vol. 47, part I, pp. 94 ff., April, 1951, part II, pp. 112 ff.

there is no control current and that the action of the control current is actually to “desaturate” the cores. It can be shown¹ that the sloping portion of the characteristic from A to B in Fig. 3.58a is identical to the back, or demagnetizing, portion of the hysteresis loop of the core material.

In order to obtain approximate equations for the gain and time constant of a self-saturated magnetic amplifier, it is convenient to approximate the empirically determined curve of Fig. 3.58a by three straight lines, as

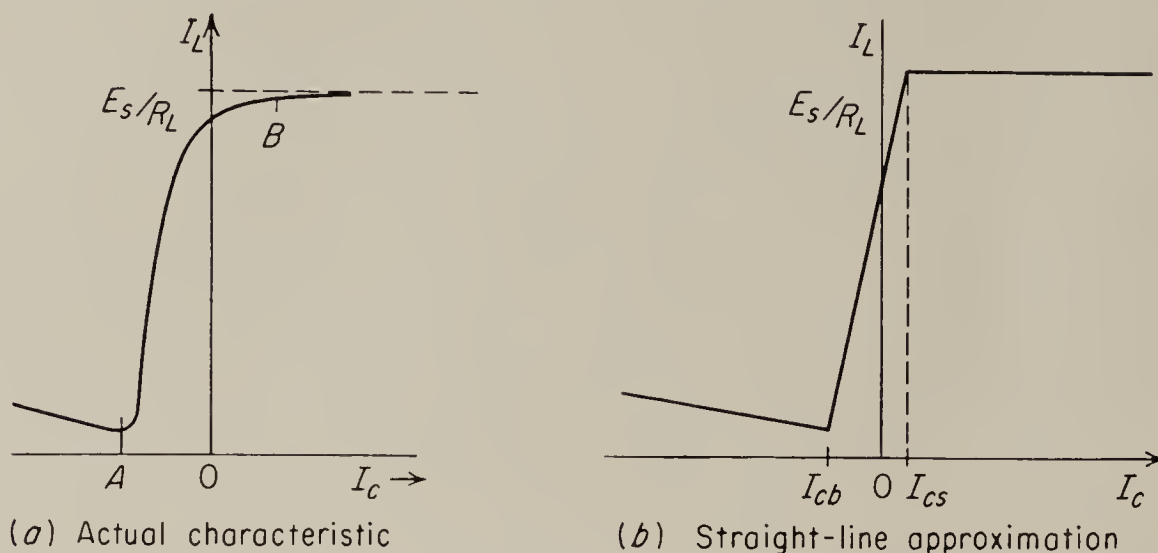


FIG. 3.58. Control characteristic of self-saturated amplifier.

shown in Fig. 3.58b. If this approximation is valid, we obtain by direct proportion

$$I_L = \left(\frac{E_s}{R_L} - I_{L \min} \right) \frac{N_c(I_c - I_{cb})}{N_c(I_{cs} - I_{cb})} + I_{L \min} \quad (3.129)$$

so that the current gain is given by

$$K_I = \frac{\partial I_L}{\partial I_c} = \frac{E_s}{R_L(NI)} N_c I_c \quad (3.130)$$

In these equations I_{cb} is the “bias” value of control current which results in the minimum output $I_{L \min}$, I_{cs} is the control current resulting in full output, and $NI = N_c(I_{cs} - I_{cb})$ is the number of control ampere-turns required to vary the output from maximum to minimum. Note that NI is a constant that depends only on the back slope of the hysteresis loop of the core material and becomes smaller as the quality of the core material is improved. The voltage gain is given by

$$K_e = \frac{\partial E_L}{\partial E_c} = \frac{R_L}{R_c} \frac{\partial I_L}{\partial I_c} = \frac{E_s N_c}{R_c(NI)} \quad (3.131)$$

Thus for a given amplifier operating at a fixed supply voltage E_s , the voltage gain is a constant and is independent of the load, whereas the current gain is inversely proportional to the load resistance. Self-saturation therefore is seen to change the magnetic amplifier from an adjustable con-

¹ *Ibid.*

stant-current generator [cf. Eq. (3.81)] to an adjustable constant-voltage generator.

From Eqs. (3.130) and (3.131) the power gain for direct current is given by

$$K_p = \frac{E_s^2 N_c^2}{R_L R_c (NI)^2} \quad (3.132)$$

The time constant of the self-saturated amplifier can be obtained by the type of argument employed in the previous analysis, since the relationship between ϕ_d and E_L derived in Eq. (3.109) is not affected by the presence of the rectifiers in the gate circuit. Thus, from Eq. (3.110)

$$L_c = N_c \frac{d\phi_d}{di_c} = \frac{K_e R_c N_c}{2N_g f} \quad (3.110a)$$

we obtain by substitution of Eq. (3.131) for K_e

$$T_c = \frac{L_c}{R_c} = \frac{E_s N_c^2}{2R_c (NI) N_g f} \quad (3.133)$$

This represents the over-all time constant for rising transients when the rectifiers are blocked. For falling transients, where circulating paths may exist through the rectifiers (depending on the exact circuit used), the time constant will be increased as discussed previously [cf. Eq. (3.108)] and may be written in general:

$$T = \frac{E_s}{2N_g f R_c (NI)} \left(\frac{N_c^2}{R_c} + \frac{N_g^2}{R_g} \right) \quad (3.134)$$

where R_g is the resistance of the circulating path in the gate winding.

3.24. The Ramey Magnetic Amplifier. A magnetic amplifier operating on a somewhat different principle from the ones described above has been described by Ramey.¹ A circuit diagram for a single-ended amplifier of this type is shown in Fig. 3.59. Note that rectifiers are employed both in the control and load winding, and that on the control side an a-c bias voltage is inserted in addition to the control signal e_c .

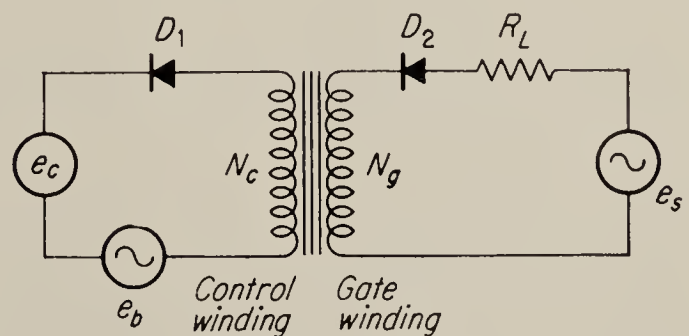


FIG. 3.59. The Ramey magnetic amplifier-basic circuit.

The operation of the amplifier may be explained most easily by assuming that e_c is a d-c voltage. The direction of this voltage is such as to apply reverse bias to the rectifier D_1 . The a-c bias voltage e_b is in phase

¹ R. A. Ramey, On the Mechanics of Magnetic Amplifier Operation, *Trans. AIEE*, vol. 70, part II, pp. 1214–1223, 1951.

with the supply voltage e_s . Thus, if $e_b = E_b \sin \omega t$, $e_s = E_s \sin \omega t$, where E_b and E_s are maximum values. Suppose now that at time $t = 0$ the core is saturated so that the flux $\phi = \phi_s$. During the first half cycle the bias voltage e_b is in a direction opposing e_c , and the direction of e_s is such that the rectifier D_2 is biased in the reverse direction. No current flows in either winding until the instantaneous amplitude of e_b exceeds e_c . At this time the voltage difference $e_b - e_c$ is applied to the control winding, and the directions of the voltages are such that the core is demagnetized. The flux therefore decreases from the saturated value ϕ_s until the bias voltage e_b is again less than e_c . At the end of the first half cycle the flux is given by the expression

$$\phi = \phi_s - \frac{1}{N_c} \int_{\omega t = \sin^{-1} \frac{e_c}{E_b}}^{\omega t = \pi - \sin^{-1} \frac{e_c}{E_b}} (E_b \sin \omega t - e_c) dt = \phi_s - \frac{A_1}{\omega N_c} \quad (3.135)$$

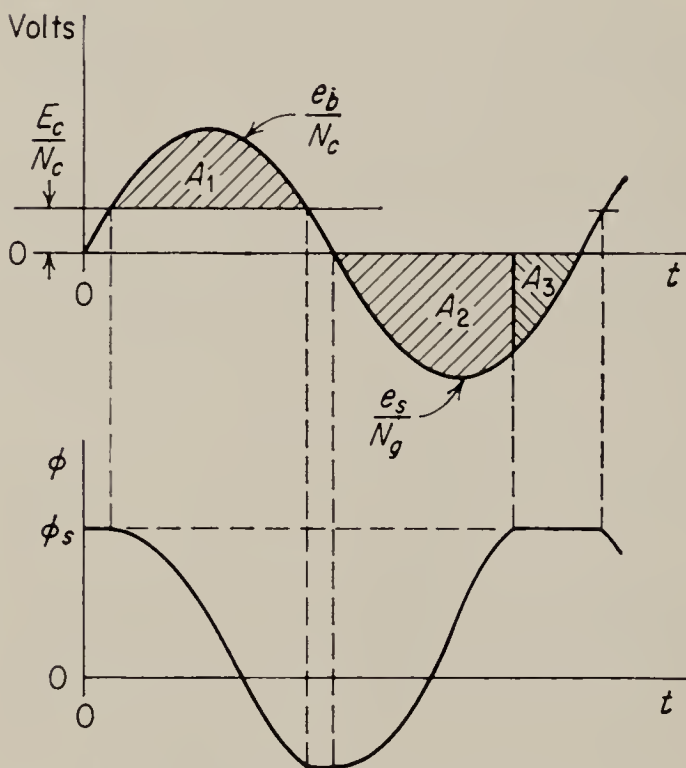


FIG. 3.60. Voltage and flux variations in the Ramey amplifier.

where A_1 is the area between the bias and control voltages as shown in Fig. 3.60. Ideally this amount of flux remains in the core until the end of the first half cycle. During the second half cycle, rectifier D_2 permits current flow in the gate circuits, but the control circuit is blocked by rectifier D_1 . If we assume that the voltage drop in the load due to the magnetizing current is negligible, the full supply voltage is applied the gate winding. The winding direction is such that the flux again increases. The flux continues to increase until the core saturates again, and at this time the supply voltage appears across

the load. The flux during the second half cycle is given by

$$\phi = \phi_s - \frac{A_1}{\omega N_c} - \frac{1}{N_g} \int_{\omega t = \pi}^{\omega t} E_m \sin \omega t dt \quad (3.136)$$

When the core again saturates $\phi = \phi_s$ and

$$\frac{A_1}{\omega N_c} = - \frac{1}{N_g} \int_{\omega t = \pi}^{\omega t = \alpha} E_m \sin \omega t dt = \frac{A_2}{\omega N_g} \quad (3.137)$$

where α is the firing angle and where A_2 is the area under the supply voltage as shown in Fig. 3.60. The output voltage consists of pulses, and its average value is proportional to the area shown in Fig. 3.60 as A_3 . It

should be clear that although a d-c control signal has been assumed for this analysis, an a-c signal at supply frequency will work just as well.

It will be noted that the amplifier operates on a *shared-time principle*. During the first half cycle the control voltage charges the core, and during the second half cycle the supply voltage discharges it. For this reason the response time of this amplifier is less than one cycle, and it is therefore much faster than conventional amplifiers.

A disadvantage of the amplifier is that the control signal must be able to furnish sufficient power to magnetize the core completely in one half of a cycle of the supply voltage. Thus the input power required tends to be larger than for the self-saturated amplifier described in Sec. 3.23, in which charging of the core takes several cycles. With modern high-grade cores the required magnetizing current is, however, quite small, and typical

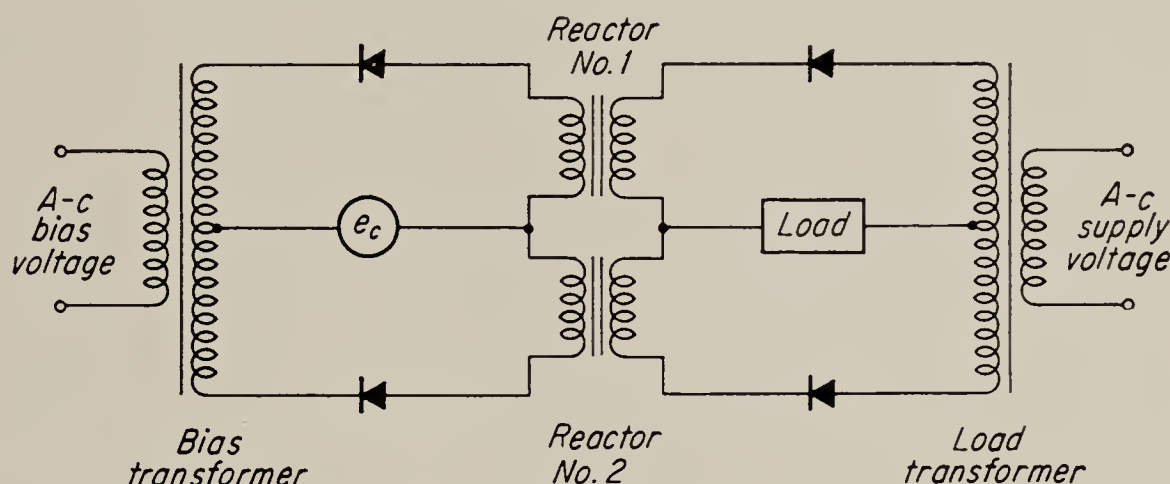


FIG. 3.61. Bridge-type Ramey circuit for full-wave output.

amplifiers of this type may have an input impedance greater than 10,000 ohms.

The basic circuit shown in Fig. 3.59 can produce only a half-wave, unidirectional load current. A somewhat more complex circuit using two reactor cores is shown in Fig. 3.61. As shown, this circuit produces a full-wave, unidirectional current in the load, but by reversing the rectifiers in the lower section the circuit is converted to furnish a half-wave, reversible current to the load.

It has been the purpose of these sections to introduce the reader to the analytic assumptions usually made in the treatment of magnetic amplifiers and to derive by their use the gains and time constants of several simple representative types. This will enable the reader to assess the operation of these devices in a control system. For design procedures and information on problems of cascading, inductive loads, and the like, the reader is referred to the texts by Storm¹ and Geyger² and to the current periodical literature.

¹ Storm, "Magnetic Amplifiers," John Wiley & Sons, Inc., New York, 1955.

² Geyger, "Magnetic Amplifier Circuits," McGraw-Hill Book Company, Inc., New York, 1954.

PROBLEMS

3.1. A half-wave thyatron amplifier is used to control the current in a pure-resistance load of 25 ohms. The supply voltage is 220 volts, and the tube drop is negligible. (a) Obtain the "small-signal" gain $\partial I / \partial \phi_a$ for $\phi_a = 90^\circ$. (b) The firing angle of the circuit is controlled by use of combined a-c and d-c bias with the a-c bias shifted 90° from the a-c plate supply. If the a-c bias is 55 volts rms and if the d-c signal voltage is E_{in} , find the over-all "small-signal" gain $\partial I / \partial E_{in}$ for an operating point defined by $\phi_a = 90^\circ$. Assume that the critical firing potential is very small compared to 50 volts.

3.2. A half-wave thyatron amplifier of the form shown in Fig. 3.15 is used to control the current in a generator field. The field has a resistance of 50 ohms and an inductance of 0.5 henry, and the thyatron amplifier is connected to a 220-volt, 60-cps single-phase supply. It is desired to vary the field current from 0.5 to 1 amp. (a) Find the range of firing angles required. (b) Find the approximate gain $\Delta I / \Delta \phi_a$ at the mid-point of the required current range, i.e., for a current of 0.75 amp. Assume that the tube drop is negligible.

3.3. The firing angle for the thyatron of Prob. 3.2 is controlled by use of combined a-c and d-c bias, with the a-c bias shifted 90° from the a-c plate supply. If the a-c bias is 110 volts rms, find the over-all gain of the circuit, $\partial I / \partial E_s$, where E_s is the d-c signal voltage, for the case $I = 0.75$ amp. Assume very small critical grid potential.

3.4. Find the average load current in the half-wave thyatron circuit of Prob. 3.2 if the load resistance is 100 ohms, if the inductance is 0.1 henry, and if the firing angle is 45° . Assume that Eq. (3.29) applies. The tube drop is 10 volts.

3.5. The load of the full-wave thyatron circuit of Fig. 3.20 consists of a 10-ohm resistance and an inductance of 0.1 henry. Plate voltage is 110 volts rms, 60 cps, and tube drop is 15 volts. (a) Find the firing angle ϕ_a for which the load current is just continuous. (b) Find the load current for firing angles $\phi_a = 30^\circ$ and $\phi_a = 150^\circ$.

3.6. In connection with the thyatron-controlled motor, show that the change-over from mode 1 to mode 2 operation and from mode 2 to mode 3 operation is not abrupt.

3.7. A half-wave thyatron circuit used to drive a d-c motor has a supply voltage of 220 volts, 60 cps. The thyatron tube drop is negligible. One of two motors is used with this circuit. One of them has an armature resistance of 0.6 ohm, an inductance of 0.001, and a K_t of 1 ft-lb/amp, while the other has a resistance of 0.7 ohm, negligible inductance, and a K_t of 0.9 ft-lb/amp. In all other respects the motors are identical. Find the ratio of maximum stall torques available from the two motors.

3.8. A relay amplifier with feedback of the sort shown schematically in Fig. 3.35 uses relays which close when the voltage across their coils exceeds 40 volts; they open when the voltage drops to 25 volts. The amplifier used between the RC filter and the relay coils has a gain of 20, and the time constant of the filter is 0.1 sec. For a 10-volt peak input signal, find the maximum frequency to which the amplifier can respond.

3.9. A series-connected magnetic amplifier like that shown in Fig. 3.45a has 5,000 turns on its control winding and 2,000 turns on its gate winding. The control resistance is 10 ohms, and the frequency of the supply is 60 cps. A load resistance of 100 ohms is connected to the amplifier, and the control voltage is adjusted to produce an average load current of 0.5 amp. At time $t = 0$, a 400-ohm resistance is shunted across the 100-ohm load. (a) Find the control current for $t < 0$. (b) Find the load and control currents at the instant after the 400-ohm resistance is added. (c) Sketch the behavior of the average load current as a function of time after $t = 0$. Neglect the pulsed nature of the output. (d) Draw an equivalent circuit that, on the average, behaves like this magnetic-amplifier circuit.

CHAPTER 4

D-C MACHINES

4.1. Introduction. D-c machines are used in the power-output stage of a large class of electromechanical servomechanisms. D-c machines are preferred over a-c machines in high-power applications because of the ease of control of the speed and direction of rotation of large d-c motors.

A widely used arrangement of d-c machines in servomechanisms is the Ward-Leonard system, in which a d-c generator drives a d-c motor. A common form of this system is shown in Fig. 4.1. The generator is driven at an essentially constant speed by a prime mover, and the field of the motor is separately excited by a constant-voltage source. (In low-power servos, the motor field may be established by a permanent magnet.) In the figure, the generator field is shown with a center tap and is driven by a

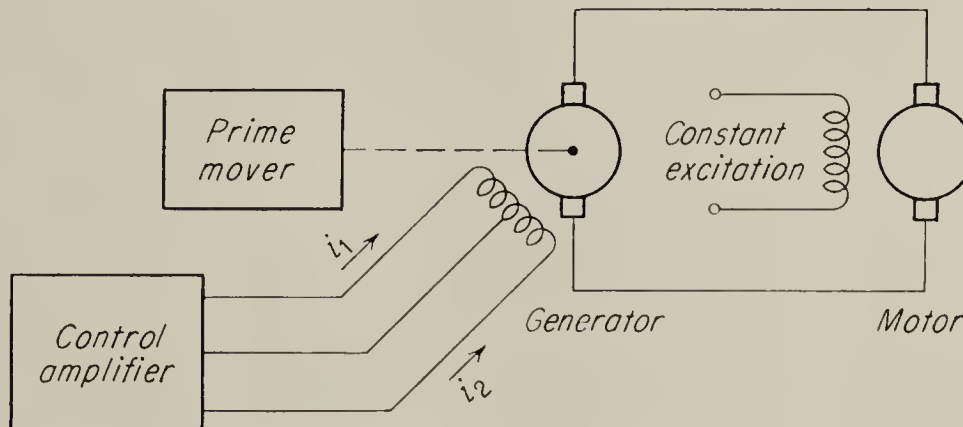


FIG. 4.1. A common form of Ward-Leonard system.

push-pull amplifier. The currents flowing in the two halves of the field winding are therefore in opposition, and their effects tend to cancel; hence the mmf produced in the air gap of the generator is proportional to the difference of the two currents. Although this method of exciting the generator field is by no means the only one used, it has the advantage of permitting reversal of the mmf, and therefore reversal of the d-c motor, without the need of actually reversing the field currents. It is therefore almost universally used when the generator field is driven by an electronic control amplifier.

The generator used in the Ward-Leonard system may be any one of several types. In low-power applications a simple d-c generator may be used, but for power levels above a few hundred watts the power ampli-

fication afforded by the ordinary generator is no longer sufficient to raise the low-power level of the electronic amplifier up to the level required by the motor. In that case one of the more elaborate, two-stage generators, such as the Rototrol, Regulex, or Amplidyne, is used. These generators will be described in detail in succeeding paragraphs.

4.2. Generated Voltage in a Simple D-C Generator. The basic relation for the open-circuit voltage produced by a d-c generator is derived from Faraday's law and may be written in the form¹

$$E_g = \frac{N_a P \phi \Omega}{p} \quad (4.1)$$

where E_g = generated voltage, volts

N_a = total number of armature conductors

P = number of field poles

ϕ = total magnetic flux per pole, webers

Ω = speed of rotation, radians/sec

p = number of parallel paths in armature winding

For a particular machine, N_a , P , and p are constants, and therefore Eq. (4.1) may be written in the simpler form

$$E_g = k_1 \phi \Omega \quad (4.2)$$

The equation indicates that, if ϕ is held constant, the generated voltage is directly proportional to the speed Ω . This principle is used in the d-c *tachometer* (also called *rate generator*). Such a machine is shown in Fig. 4.2 and is seen to be a small generator, usually built with a stable, permanent magnet to provide the field excitation. The commutator of a d-c-tachometer armature usually has a relatively large number of segments so as to provide as smooth an output as possible. Nevertheless the output contains a ripple component, commonly of the order of 5 to 10 per cent of the d-c voltage; this sometimes causes difficulty in very high performance loops. Equation (4.2) indicates that the d-c output voltage is accurately proportional to the speed, with no time lag; however, at very high speeds, brush bounce has the effect of reducing the accuracy and at the same time increasing the ripple and noise output. If calibration accuracy is an important requirement, care must be taken never to short-circuit or load the tachometer heavily, since the resulting mmf of armature reaction² may demagnetize the field magnets to some extent and thus change the calibration.

¹ See, for instance, Bull, "DC Machinery," John Wiley & Sons, Inc., New York, 1947, p. 46.

² *Ibid.*, chap. 11. Also Dawes, "A Course in Electrical Engineering," vol. 1, "Direct Currents," 4th ed., McGraw-Hill Book Company, Inc., New York, 1952, pp. 415ff.

When a generator is used for power amplification, as in the Ward-Leonard system, it is driven at constant speed. Hence Ω in Eq. (4.2) is a constant that may be combined with k_1 , so that only ϕ and E_g are variable, ϕ being the independent variable. When the generator is open-circuited, ϕ depends only on the field current.

The relationship between the flux and the field current is ordinarily expressed graphically in the form of a hysteresis loop or *saturation*

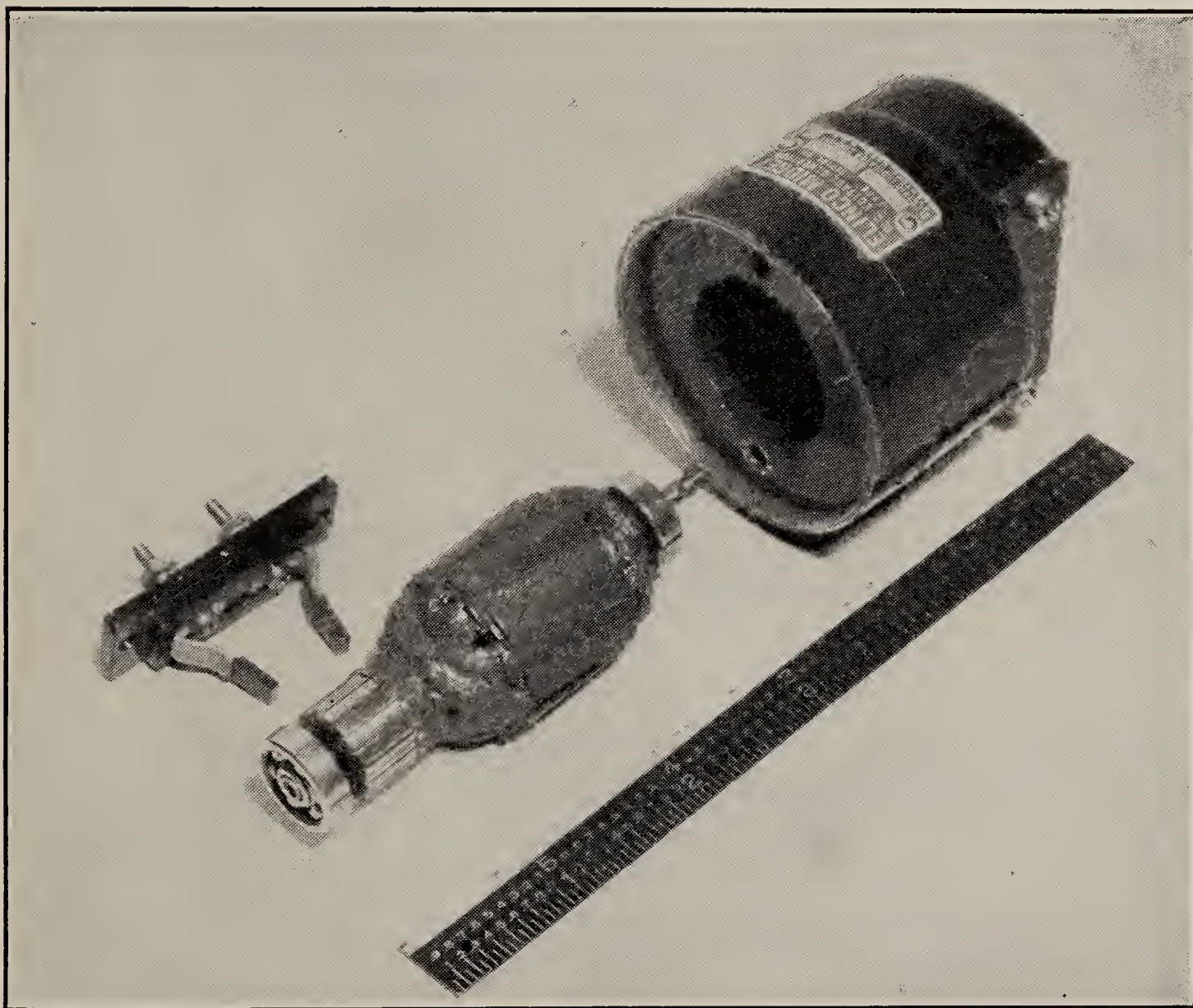


FIG. 4.2. Stator, armature, and brush rigging of a typical d-c tachometer.

curve of the magnetic path in which the flux is established. In most designs for generators, a steel of low retentivity is used. The hysteresis loop is thus narrow enough for the hysteresis effect to be ignored, at least in a first approximation, and an “average” saturation curve is used to describe the characteristics of the magnetic path (see Fig. 4.3). A further simplification is possible in many cases by virtue of the fact that in normal operation the voltage of the generator will be near zero,¹ where the

¹ This statement is true for position servos, where the motor normally does not turn or turns only slowly. In speed-control systems, where the motor normally runs at a relatively high speed, a linear approximation to the magnetization curve can also be made; however the slope of the straight-line approximation then depends on the operating point.

magnetization curve is approximately a straight line. In this case we may say that the flux is linearly proportional to the ampere-turns, or

$$\phi = \frac{N_f i_f}{\mathcal{R}} \quad (4.3)$$

where N_f is the number of turns of the field and \mathcal{R} is the reluctance of the magnetic path expressed in suitable units. A combination of Eqs. (4.2)

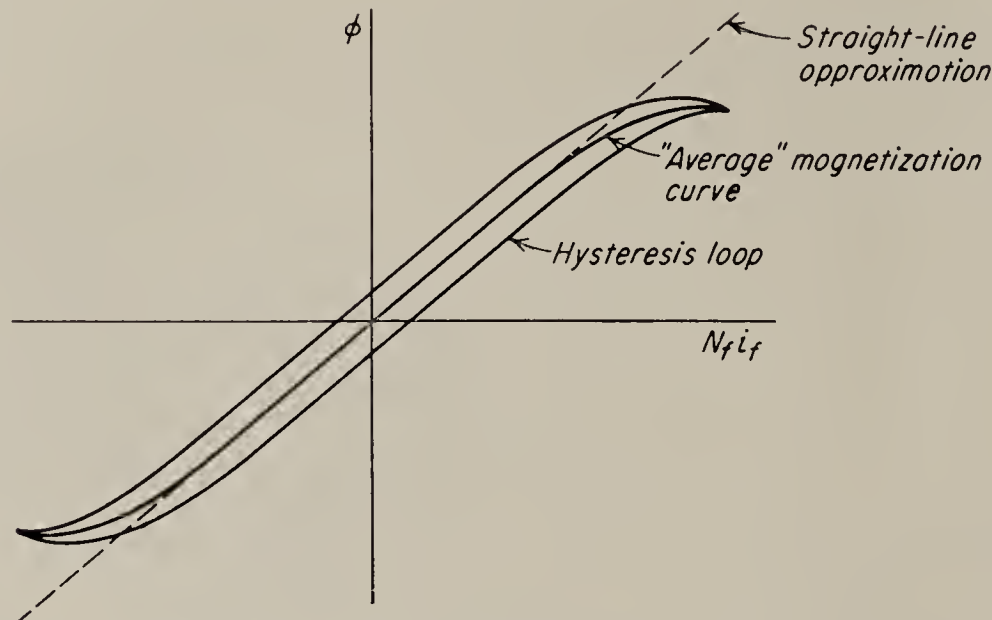


FIG. 4.3. Hysteresis loop and "average saturation curve" of a generator.

and (4.3) results then in a linear relation between field current and generated voltage:

$$E_g = k_g i_f \quad (4.4)$$

The field circuits of almost all d-c generators are highly inductive, since a large number of turns are used to obtain the required magnetomotive force with a minimum of field current. Hence, the inductance must be considered, as in the circuit shown in Fig. 4.4. We have

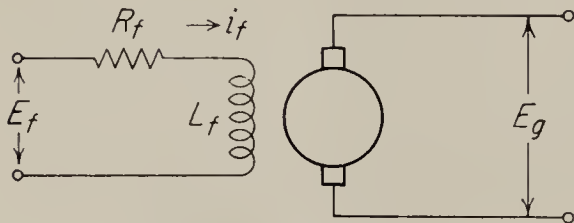


FIG. 4.4. The equivalent circuit of a d-c generator.

$$E_f = R_f i_f + L_f \frac{di}{dt} \quad (4.5)$$

if the armature current is zero or constant. By use of Laplace transforms we can solve for the field current

$$\hat{i}_f = \frac{\hat{E}_f}{R_f + L_f s} = \frac{(1/R_f) \hat{E}_f}{1 + T_f s} \quad (4.6)$$

where $T_f = L_f/R_f$ is the field time constant. In this equation the resistance R_f is the total resistance of the field circuit together with the output resistance of the driving source considered as a Thévenin generator;¹ E_f is the Thévenin voltage of the source.

¹ Everitt and Anner, "Communication Engineering," 3d ed., McGraw-Hill Book Company, Inc., 1956, p. 119.

When a center-tapped field is used with a push-pull circuit, as shown in Fig. 4.1, an equation very similar to Eq. (4.6) can be derived. If the push-pull amplifier stage consists of two vacuum tubes with an amplification factor μ and a plate resistance r_p and if the voltages applied to the grids of these tubes are e_g and $-e_g$, respectively, then an equivalent circuit for the push-pull arrangement takes the form of Fig. 4.5. Here the arrows indicate the assumed positive direction of the voltages and currents. With the current assumed as shown, the effective mmf is $N_f(i_{f1} - i_{f2})$, if both halves of the field winding have the same number of turns, N_f . R'_f represents the resistance of each field winding alone; M is the mutual inductance between the two halves; E_{bb} and Z_b are the voltage and output impedance, respectively, of the amplifier power supply.

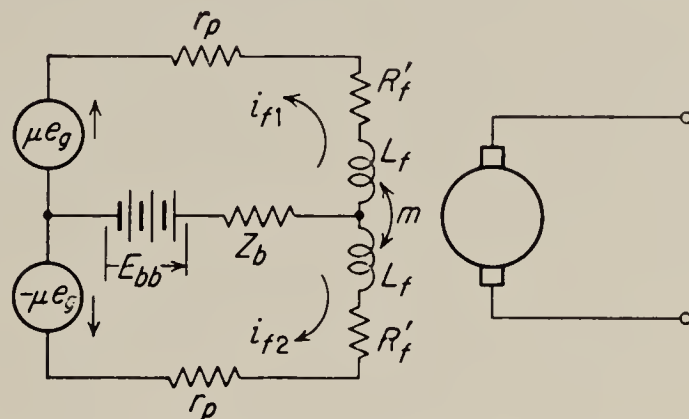


FIG. 4.5. Equivalent circuit of a d-c generator with push-pull field.

The mmf produced by the two fields may be found as a function of the control voltage e_g by use of Kirchhoff's laws. In writing the mesh equations for the circuit, it is convenient to let

$$\begin{aligned} Z &= r_p + R'_f + L_f s + Z_b \\ Z_M &= Z_b - M s \end{aligned} \quad (4.7)$$

and

Then the two mesh equations take the form

$$\begin{aligned} -\mu \hat{e}_g + \hat{E}_{bb} &= Z \hat{i}_{f1} + Z_M \hat{i}_{f2} \\ \mu \hat{e}_g + \hat{E}_{bb} &= Z_M \hat{i}_{f1} + Z \hat{i}_{f2} \end{aligned} \quad (4.8)$$

$$\text{whence} \quad \hat{i}_{f1} - \hat{i}_{f2} = \frac{-2\mu \hat{e}_g}{Z - Z_M} = \frac{-2\mu \hat{e}_g}{r_p + R'_f + (L_f + M)s} \quad (4.9)$$

where in the second step we substitute for Z and Z_M from Eq. (4.7). The mmf is therefore

$$\mathfrak{F} = N_f(\hat{i}_{f1} - \hat{i}_{f2}) = \frac{-2N_f\mu \hat{e}_g}{r_p + R'_f + (L_f + M)s} = \frac{-2N_f\mu \hat{e}_g}{(r_p + R'_f)(1 + T_f s)} \quad (4.10)$$

where $T_f = (L_f + M)/(r_p + R'_f)$. If $L_f + M$ and $r_p + R'_f$ are thought of as the equivalent inductance and resistance, respectively, of the push-pull field winding and if we let $-2\mu \hat{e}_g$ be equal to the driving voltage \hat{E}_f of the circuit of Fig. 4.4, this result is seen to be identical with the one obtained in Eq. (4.6). Hence Eqs. (4.2), (4.3), and (4.6) may be combined to obtain the expression for the open-circuit voltage:

$$\hat{E}_g = \frac{k_1 \Omega k \phi N_f / R_f}{T_f s + 1} \hat{E}_f = \frac{(k_g / R_f) \hat{E}_f}{T_f s + 1} \quad (4.11)$$

In later discussions we shall make use of the equivalence of push-pull and single-ended field coils and ordinarily carry through the various analyses on the basis of a single-ended field, even though the push-pull type is more common.

4.3. The Effects of Hysteresis and Eddy Currents. In the previous discussion the properties of the iron making up the magnetic paths of the generator were not considered. However, in practice, the imperfections in the iron will affect the generator performance. The fact that the iron saturates results in a fundamental nonlinearity between generated voltage and applied voltage. Once the generated voltage reaches the saturation

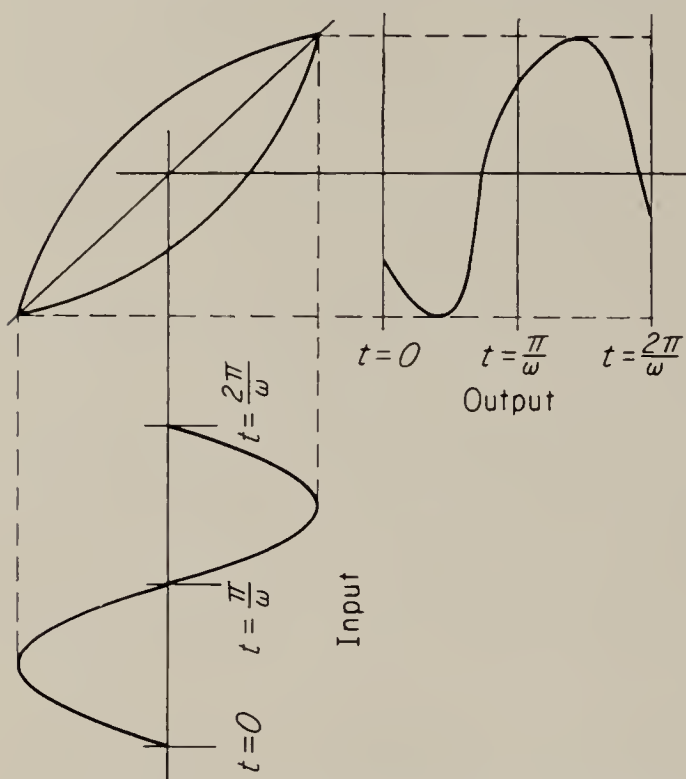


FIG. 4.6. Phase lag produced by hysteresis.



FIG. 4.7. Gain change produced by hysteresis.

value, the steady-state gain of the machine (defined as $\partial E_g / \partial E_f$) decreases rapidly and approaches zero for very large values of E_f .

Hysteresis of the iron causes a phase lag between input and output of the generator that is a function of the amplitude of the input but is independent of frequency. This effect is best demonstrated by assuming that the input voltage is sinusoidal and plotting the resulting output voltage through one complete cycle as in Fig. 4.6. This construction also indicates that the output wave contains harmonic components not present in the input. The extra phase lag of the fundamental component, which is not predicted by the linear approximation, may result in poorer stability of the system of which the generator is a part. Furthermore, comparison of the hysteresis loops obtained on a given piece of iron for various amplitudes of excitation indicates that the average slope for small amplitudes, and therefore the gain, is less than that for larger amplitudes (Fig. 4.7). Thus the combined effect of hysteresis and saturation on the gain is to reduce the gain, for both very small and very large amplitudes, compared

to its value at intermediate amplitudes. A complete mathematical description of all of these effects results in nonlinear differential equations of considerable complexity, which cannot, at the present time, be solved completely. Hence many of these effects are often neglected in practical analyses, for the sake of simplicity.

An additional imperfection of the iron is the generation of eddy currents in the field poles whenever the field current changes. This effect may be considered to be equivalent to having a large number of closed circuits in the iron mutually coupled to the field circuit and to each other. Since these circuits must be chosen somewhat arbitrarily, a simplifying assumption is made that yields an approximate answer sufficiently accurate for many purposes. The multiplicity of mutually coupled circuits is replaced by an equivalent single eddy-current circuit that is mutually coupled to the field winding. This situation is diagrammed in Fig. 4.8, where L_e and R_e represent the effective inductance and resistance, respectively, of the eddy-current path, and where M is the mutual inductance between the eddy current and field circuits. If the number of turns of the field winding is N_f and if the number of effective turns for the eddy-current circuit is N_e , then the following three simultaneous equations (written in Laplace transform notation) describe this situation:

$$\hat{E}_f = (R_f + L_f s)\hat{i}_f - M s\hat{i}_e \quad (4.12)$$

$$0 = -M s\hat{i}_f + (R_e + L_e s)\hat{i}_e \quad (4.13)$$

$$\hat{E}_g = k(N_f\hat{i}_f - N_e\hat{i}_e) \quad (4.14)$$

The signs of M in Eqs. (4.12) and (4.13) must be the same but can be either positive or negative, depending on the sense assumed for the winding of the coil representing the eddy-current circuit. The sign of N_e in Eq. (4.14) must, however, correspond to the signs chosen for M in such a way that the effect of the eddy current i_e is to oppose the field current i_f . This is required by Lenz's law. Simultaneous solution of the three equations then yields

$$\hat{E}_g = \frac{k\hat{E}_f \left(\frac{N_f}{R_f} + \frac{N_f L_e}{R_f R_e} s - \frac{N_e M}{R_f R_e} s \right)}{\frac{L_f L_e - M^2}{R_f R_e} s^2 + \left(\frac{L_f}{R_f} + \frac{L_e}{R_e} \right) s + 1} \quad (4.15)$$

This equation can be simplified if we assume that the magnetic paths for both the eddy-current circuit and the field winding are identical. With

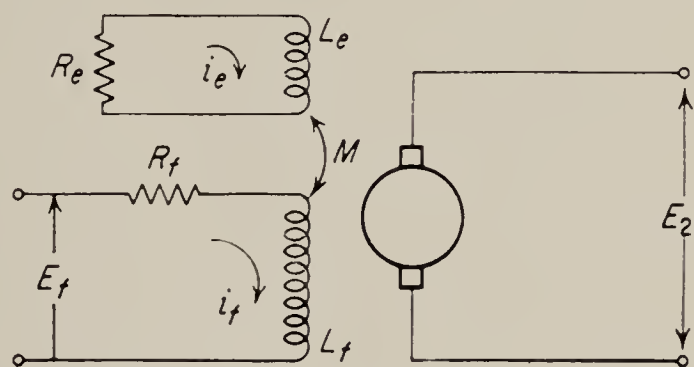


FIG. 4.8. Equivalent circuit accounting for eddy currents.

this assumption we have

$$L_e = \frac{N_e^2}{N_f^2} L_f \quad (4.16)$$

This assumption also implies perfect coupling of the two circuits, so that

$$M = \sqrt{L_e L_f} = L_f \frac{N_e}{N_f} \quad (4.17)$$

Substitution of these relations in Eq. (4.15) gives

$$\hat{E}_g = \frac{(kN_f/R_f)\hat{E}_f}{[(L_f/R_f) + (L_e/R_e)]s + 1} = \frac{(kN_f/R_f)\hat{E}_f}{(T_f + T_e)s + 1} \quad (4.18)$$

where T_e is the “eddy-current-circuit” time constant. We note that, at least to the accuracy of the assumptions made in this development, the effect of eddy currents is to increase the time constant of the generator. This additional time lag may become particularly serious when, in order to obtain rapid response, R_f is made large. A pentode amplifier driving the generator field has this effect. Equation (4.18) shows that, even if T_f can be made negligibly small, the time constant T_e will still cause a lag in the transfer function relating E_f to E_g . It is clear that the suppression of eddy currents in the field of a generator is of great importance where maximum response speed is desired; and for this reason extra-thin punchings are used in the laminations of the field poles of generators designed for service in servo control systems.

4.4. Factors Affecting Terminal Voltage of a Loaded Generator. Up to this point we have been concerned primarily with the open-circuit voltage produced by a generator and its relation to the voltage applied to the field. The implicit justification for this preoccupation with open-circuit voltage has been Thévenin’s theorem.¹ This theorem, which strictly applies only to linear systems, may be extended to apply even to some nonlinear systems by supposing the output impedance to be in part a function of the load current. The knowledge of the transfer function from generator input to open-circuit voltage is not a complete description of the generator under load; the output impedance is also required.

When a generator delivers current to a load, the output voltage drops because of the effects of the resistance and inductance of the armature winding, armature reaction, and mutual inductance between armature and field circuits. The armature resistance usually includes the brush resistance and is therefore not completely independent of the armature current. Nevertheless, a linear relation between armature resistance drop and armature current is almost always assumed, because in most

¹ Everitt, *loc. cit.*

generators the voltage drop due to the nonlinear brush contact is quite small compared to the total voltage.

Armature reaction, defined as the mmf set up by the armature current, may decrease the air-gap flux and thereby reduce the output voltage. The decrease of air-gap flux may be due either to a demagnetizing component of the armature reaction or to a distortion of the air-gap flux pattern.¹ Since the mmf of armature reaction is directly proportional to the armature current, it may be assumed that the reduction in flux, and therefore the reduction in output voltage, is proportional to the current. This means that the armature-reaction effect is similar to the resistance effect, and the two may therefore be lumped together as an “effective” resistance for purposes of analysis. This “effective” resistance cannot, of course, be measured by simple voltmeter-ammeter tests but must be determined by some form of load or short-circuit test.

Since armature reaction also causes sparking at the brushes,² considerable effort is often expended to reduce it. This is particularly true in servo generators, which are often called upon to deliver heavy accelerating or braking currents, and which would spark badly if armature reaction were permitted to go unchecked. Both commutating poles and pole-face windings³ are therefore used extensively in these generators to improve commutation. When such compensation of armature reaction is used to reduce sparking, it will also diminish the other effects of armature reaction. Hence armature reaction is often neglected altogether in the analysis of servo generators. However, in certain types of generators, particularly in Amplidynes, armature reaction is essential to the operation.

The self-inductance of the armature circuit includes the inductance of the armature winding together with any series windings used on commutating poles and pole-face compensating windings. If compensating windings are used, the mutual inductance between the compensating winding and the armature winding subtracts from the total inductance. Thus, if L_a is the inductance of the armature, L_c the inductance of the compensating winding, and M the mutual inductance between them, then the effective inductance⁴ becomes $L_a + L_c - 2M$. Hence, a compensating winding, in addition to reducing armature reaction, also reduces the inductance. This is desirable, since inductance causes a time lag between the open-circuit voltage and the load current.

Mutual inductance between armature and field circuits is important only in compound generators using a series field, since otherwise the flux

¹ Dawes, *loc. cit.*

² *Ibid.*, pp. 426–428.

³ *Ibid.*

⁴ This equation is derived in a number of elementary texts; see, for instance, Dawes, *ibid.*, p. 319.

paths of the armature and field circuits are essentially at right angles to each other and the coupling is very small. Compound generators are not ordinarily used in servomechanisms, but we mention mutual inductance here in anticipation of the discussion of the more elaborate generators such as Rototrols and Amplidynes, where the effect is not always negligible.

In order to assess the general importance of mutual inductance in d-c generators, we compute the relative magnitudes of the voltages induced

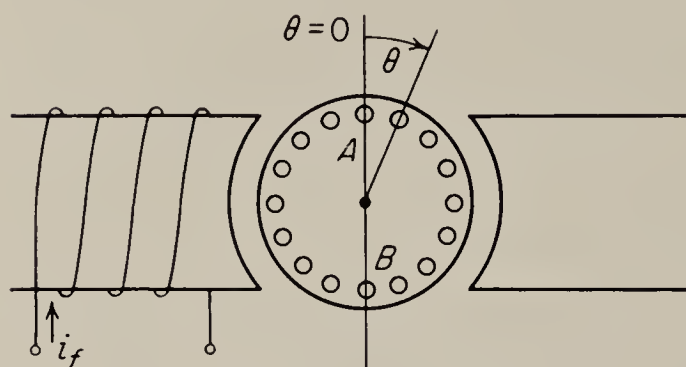


FIG. 4.9. Simplified schematic of armature winding.

in the armature by speed action and those induced by mutual inductance. For this purpose, consider an armature winding rotating in an alternating field, as shown in Fig. 4.9. The armature voltage generated in each turn of the armature winding is given by

$$e_g = \frac{d\phi}{dt} \quad (4.19)$$

where ϕ is the flux in webers linking the particular turn in question. The flux threading a particular turn is a periodic function of the space position of that turn. Thus the turn made up of the conductors *A* and *B* is linked by the maximum amount of flux for the armature position shown in the figure. This position is defined as $\theta = 0$. As the armature rotates, the flux decreases to zero as θ becomes 90° , reverses sign (relative to the turn under discussion), reaches a maximum for $\theta = 180^\circ$, goes to zero again, and again reaches a maximum when $\theta = 360^\circ$. With the usual shape of the field poles the function of flux versus θ is a flat-topped wave, but for purposes of analysis it is convenient to assume a sinusoidal wave. If the flux at any particular position of the armature also varies as a sinusoidal function of time, we may write

$$\phi = \Phi_m \cos \theta \cos \omega t = k_\phi I_{f \max} \cos \omega t \cos \theta \quad (4.20)$$

where ω and $I_{f \max}$ are the frequency and amplitude, respectively, of the field current and k_ϕ is a constant, linearity of the magnetic path being assumed here. Hence, the differentiation indicated in (4.19) gives

$$e_g = -k_\phi I_{f \max} \left(\omega \cos \theta \sin \omega t + \frac{d\theta}{dt} \cos \omega t \sin \theta \right) \quad (4.21)$$

The effective value of e_g may then be written in the standard complex notation

$$E_g = -\Omega k_\phi I_f \sin \theta + j\omega k_\phi I_f \cos \theta \quad (4.22)$$

where we have substituted the speed of rotation, Ω , for $d\theta/dt$ in accordance with the nomenclature used in Eq. (4.1). Note that the first term on the right is proportional to speed (Ω) and independent of frequency. This term represents the voltage induced by speed action and may be designated as E_Ω . It is a maximum in those armature turns for which $\theta = 90^\circ$. The second term on the right is independent of speed but proportional to frequency and is maximum in those turns for which $\theta = 0^\circ$. This is the voltage due to the effect of mutual inductance and is designated as E_M . Inspection of Eq. (4.22) reveals that

$$E_{M \max} = \frac{\omega}{\Omega} E_{\Omega \max} \quad (4.23)$$

i.e., the ratio of mutually induced voltage to voltage generated by speed action is equal to the ratio of the frequency of the field current to the speed of rotation of the armature, both expressed in radians per second. (In a generator with more than two poles, Ω is the speed in *electrical* radians per second, i.e., the actual speed multiplied by the number of pairs of poles.) In most servo systems the ratio ω/Ω is very much less than unity, and mutual-inductance effects are therefore usually negligible compared to speed effects.

The complete behavior of a generator under load may now be expressed in the form of the two simultaneous equations given below. These equations include all the factors discussed in this section, even though the reader may have formed the impression (correct in many cases) that the only factor not normally neglected is the armature resistance. We do not include the effect of eddy currents and hysteresis in the iron, however, since inclusion of these effects only complicates the final expressions without adding much in the way of new information.

$$\hat{E}_f = (R_f + L_f s) \hat{i}_f + M_{af} s \hat{i}_a \quad (4.24)$$

$$\hat{E}_g - (R_a + L_a s) \hat{i}_a - M_{fa} s \hat{i}_f = \hat{V}_o \quad (4.25)$$

In these equations E_f is the voltage applied to the field; L_f and R_f are the inductance and total resistance, respectively, of the field circuit; $M_{af} = M_{fa}$ is the mutual inductance between armature and field circuits; E_g is the open-circuit voltage produced by the generator; R_a is the "effective" armature resistance, including armature reaction and brush drop; L_a is the armature inductance; and V_o is the terminal voltage existing under load conditions. E_g and i_f are related by Eq. (4.4):

$$\hat{E}_g = k_g \hat{i}_f \quad (4.4)$$

Since we are interested in the output impedance of the generator, it is

convenient to rewrite Eqs. (4.24) and (4.25) in such a form that E_f and i_a are the independent variables and V_o and i_f the dependent variables:

$$\hat{E}_f - M_{af}s\hat{i}_a = (R_f + L_fs)\hat{i}_f \quad (4.26)$$

$$-(R_a + L_as)\hat{i}_a = (-k_g + M_{fa}s)\hat{i}_f + \hat{V}_o \quad (4.27)$$

Solving for \hat{V}_o , we obtain

$$\begin{aligned} \hat{V}_o &= \frac{\hat{E}_f(k_g - M_{fas})}{R_f + L_fs} - \frac{(R_f + L_fs)(R_a + L_as) - M_{af}^2s^2 + M_{af}k_gs}{R_f + L_fs}\hat{i}_a \\ &= \frac{k_g/R_f - (M_{fa}/R_f)s}{T_fs + 1} \hat{E}_f - \left(R_a + L_as + M_{af}s \frac{k_g - M_{af}s}{R_f + L_fs} \right) \hat{i}_a \end{aligned} \quad (4.28)$$

The first term in this equation represents the open-circuit voltage, and except for the very small mutual-inductance term in the numerator, this term is identical with the expression for the open-circuit voltage already derived [Eq. (4.11)]. The second term shows the effect of the load current; hence the quantity in brackets multiplying the load current represents the output impedance. It is seen that the output impedance consists primarily of $R_a + L_as$, the resistance and self-inductance of the armature. The effect of mutual inductance can be estimated approximately if we assume that $k_g \gg M_{af}s$ and $R_f \gg L_fs$. This can, of course, be true only at low frequencies; however, if these relations do hold, the mutual-inductance term reduces approximately to $M_{af}sk_g/R_f$ and has the form of an additional inductance in the armature circuit, which either subtracts from or adds to the self-inductance L_a (depending on the sign of M_{af} relative to the sign of k_g). In a noncompound generator the entire effect is usually extremely small.

4.5. Figure of Merit. A figure of merit used very commonly for all types of power-amplifying equipment is the product of power gain and bandwidth. In order to compute this figure for the generator as a function of some of the design features, it is first necessary to give precise definitions for speed of response and power gain. Both of these quantities will be defined for a generator connected to a pure resistance load; and since the power gain is a function of the value of the load resistance, the definitions are based on the use of a standard load resistance equal to the generator output resistance. The *response speed* is then defined as the frequency in radians per second of the sinusoidally varying control voltage E_f which results in a reduction of the amplitude of the load voltage by a factor of $1/\sqrt{2}$ from the very low frequency value. This is the commonly used *half-power-point* definition. The *power gain* is defined as the ratio of the power dissipated in the standard load resistance to the power supplied to the field by the control voltage, both powers to be measured at direct current or at very low frequency.

For a generator with an output impedance that is purely resistive (i.e.,

for which the effects of the mutual and self-inductances in the armature circuit are negligible), the speed of response is given by

$$\omega_c = \frac{1}{T_f + T_e} = \frac{R_f}{L_f(1 + T_e/T_f)} \quad (4.29)$$

where R_f and L_f are the resistance and inductance of the field circuit, respectively, and T_e is the eddy-current time constant (see Sec. 4.3). The power input is

$$P_i = \frac{E_f^2}{R_f} \quad (4.30)$$

and the power output is

$$P_o = \left(\frac{E_g}{R_a + R_L} \right)^2 R_L \quad (4.31)$$

where R_a is the armature resistance, including the effects of armature reaction and brush resistance, and R_L is the load resistance. If $R_L = R_a$, this becomes

$$P_o = \frac{E_g^2}{4R_a} = \frac{k_g^2 E_f^2}{4R_f^2 R_a} \quad (4.32)$$

Hence the figure of merit becomes

$$F = \frac{k_g^2}{4R_a L_f (1 + T_e/T_f)} \quad (4.33)$$

The quantities entering into this expression are still related to some extent. Thus, k_g is proportional to the number of turns on the field-coil winding [see Eqs. (4.1), (4.3), and (4.4)], while L_f is proportional to the square of the number of field turns. Hence, increasing the number of turns will not affect the figure of merit. By similar reasoning it may be shown that the number of turns of the armature, the number of poles, the type of armature winding, etc., all have a negligible effect on the figure of merit. On the other hand, since both k_g and L_f are inversely proportional to the magnetic reluctance, any decrease of reluctance will result in an over-all improvement. Such a decrease might be produced by a reduction of the air gap or the use of a higher grade of iron. The figure of merit is proportional to the square of the speed, since k_g is proportional to speed. This indicates that higher performance can be obtained from a given machine by running it at higher speed. This fact is used in some aircraft generators.

The particular figure of merit computed here gives an indication of which of two generators is "better," if the only criteria are power gain and bandwidth. However there are other criteria, and therefore other figures of merit, that may be more pertinent in a given situation, and hence no single figure of this sort is ever adequate to give a complete evaluation of

the worth of a machine in a given situation. In some cases the product of voltage gain and bandwidth is more useful, and in most cases, factors such as cost, overload capacity, size and weight, etc., are very important and must be carefully considered in an over-all design.

4.6. The Rototrol Generator. A single generator of the form described in the preceding paragraphs is capable of power gains of the order of 10 to 20. When larger power gains are required, it becomes necessary to cascade two or more generators, i.e., to use one generator to excite the field of the second, etc. The *Rototrol*¹ is essentially such a cascade arrangement of two generators, but since the first, or pilot, generator is partially self-excited, the power gain can theoretically be made infinite. A schematic diagram is shown in Fig. 4.10.

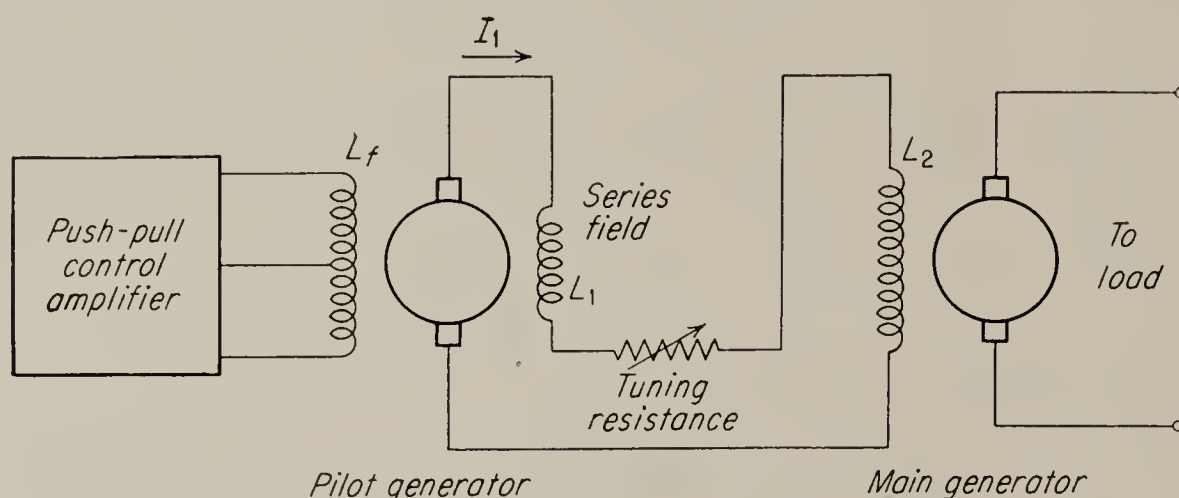


FIG. 4.10. Schematic of the Rototrol generator.

Both the pilot generator and the main generator are driven at constant speed by a prime mover not shown in the figure. The pilot generator has at least two fields, one of which is a series field L_1 and the other the control field. In the figure a pilot generator with a single push-pull control field is shown; however many pilot generators, particularly those designed for heavy-duty industrial service, may have four or five separate control fields in addition to the series field. This multiplicity of control fields makes it possible to build a control system for fairly high performance without the use of vacuum tubes. Such a system is shown in Fig. 4.13. Since the basic operation of a Rototrol is not affected by the number of control fields used, we shall concentrate on the single-control-field type in the following paragraphs and show later (Sec. 4.8) what changes are necessary in the theory when additional control fields are used.

4.7. Analysis of the Rototrol Generator. The mmf responsible for the air-gap flux in the pilot generator is the sum of the ampere-turns of the control field and the series field. For a given value of control-field current and series-field current the generated voltage of the generator may be determined from an average magnetization curve of the sort described

¹ W. H. Formhals, *Rototrol, a Versatile Electric Regulator*, *Westinghouse Engr.*, vol. 2, pp. 51–54, May, 1942.

earlier in connection with the simple generator. Such a magnetization curve is shown as the solid line of Fig. 4.11. Since the armature circuit of the pilot generator is closed through the pilot-generator series field, tuning resistance, and main generator field (see Fig. 4.10), the current and therefore the magnetomotive force of the series field is determined in the steady state by the relation

$$N_1 I_1 = \frac{N_1 E}{R} \quad (4.34)$$

E is the generated voltage of the pilot generator, as determined by the magnetization curve, and R is the total resistance of the pilot armature

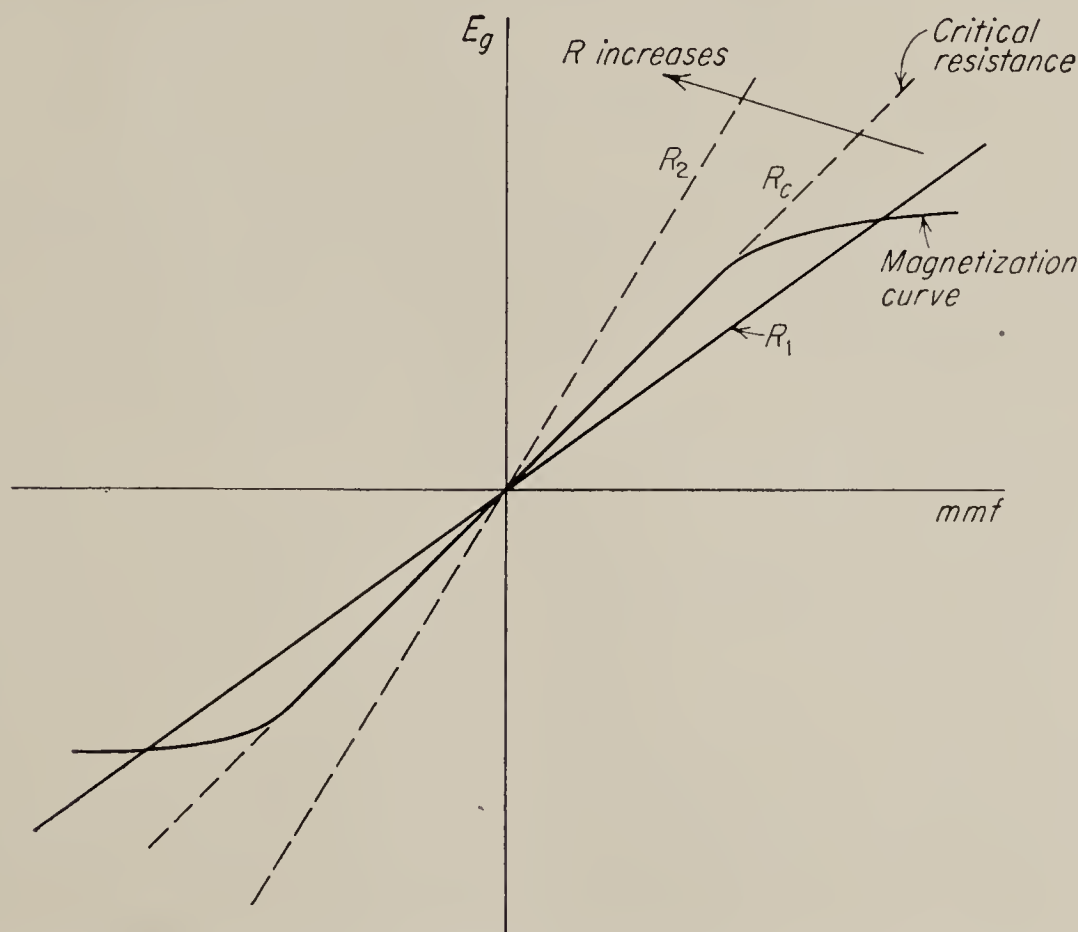


FIG. 4.11. Points of equilibrium on Rototrol operating curves.

circuit. Equation (4.34) is a linear relationship and is plotted for three representative values of R on Fig. 4.11 (the dotted lines). In the absence of control-field current the series-field magnetomotive force is the only one present; and since the generator must simultaneously satisfy the relations imposed upon it by the magnetization curve and Eq. (4.34), it can operate in equilibrium only at the points where these two relations intersect.

We see in Fig. 4.11 that, for one adjustment of the armature-circuit resistance R_c , the straight part of the magnetization curve and the resistance line coincide; hence for this resistance there are an infinite number of intersection points, any of which represents an equilibrium point for the generator. The value of resistance that achieves this adjustment is referred to as the *critical resistance*, and the Rototrol is said to be *criti-*

cally tuned when the resistance is adjusted to this value. A critically tuned Rototrol is capable of delivering a constant output voltage, even though the control-field current is zero. It should be noted that in practice it is never possible to get perfect coincidence between the resistance line and the magnetization curve, since the magnetization curve does not possess a true straight-line portion. Hence one finds that the output voltage always drifts slowly to an equilibrium point, even with the best adjustment of the resistance. This imperfection is not, however, of any great consequence in a closed feedback loop.

When the control-field current changes from zero to a finite value, the

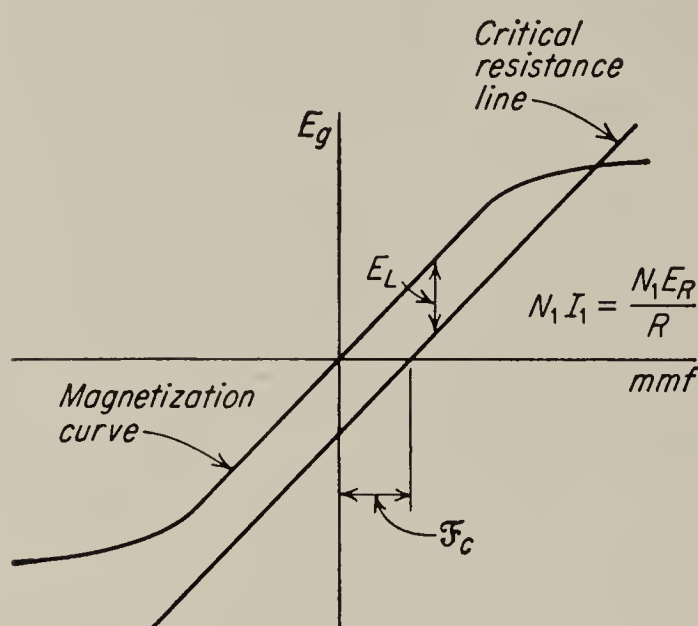


FIG. 4.12. Effect of control-field current on critical-resistance line.

machine will presumably leave its steady-state condition. Then Eq. (4.34) is no longer correct because of the inductance in the armature circuit, unless E is redefined as the resistive component of the generated voltage. If this is done, the total mmf acting on the air gap of the pilot generator may be written as

$$\mathfrak{F}_{\text{total}} = \mathfrak{F}_c + \frac{N_1 E_R}{R} \quad (4.35)$$

where \mathfrak{F}_c is the mmf supplied by the control field and E_R has been substituted for E in Eq. (4.34) to emphasize the fact that it is the resistive component. This new situation may now be diagrammed as in Fig. 4.12. The critical-resistance line is shown shifted to the right by the amount of the additional mmf supplied by the control field (\mathfrak{F}_c). The straight-line part of the magnetization curve is parallel to the critical-resistance line, and there is a fixed voltage difference E_L between them. For any assigned mmf the voltage read from the magnetization curve is E_g , the generated voltage. The voltage read from the resistance line is equal to the total resistance drop in the armature. The difference E_L must therefore be the voltage drop due to the total inductance of the armature circuit. The rate of change of the armature current is directly proportional to this voltage and is constant as long as the resistance line and magnetization curve are parallel. E_L is proportional to \mathfrak{F}_c in the linear region of the magnetization curve; therefore the rate of change of main-generator field current and the output voltage are proportional to the mmf of the control field. In other words, the output voltage is proportional to the integral of the control-field current; hence a critically tuned Rototrol acts as an integrating device.

It is now a relatively simple matter to derive the transfer function of

the Rototrol. If the magnetization curve of the pilot generator is taken to be a straight line, the internally generated voltage of the pilot generator is

$$\hat{E}_g = k_g \left(\hat{i}_f + \frac{N_1}{N_f} \hat{i}_1 \right) \quad (4.36)$$

where k_g is the generator gain constant, as defined in Eq. (4.4), N_1 is the number of turns on the series field, and N_f is the effective number of control-field turns. This expression ignores mutual inductance between the two fields and the armature and also the effect of eddy currents in the iron. The armature current i_1 is now given by

$$\hat{i}_1 = \frac{\hat{E}_g}{R + Ls} \quad (4.37)$$

where R is the total resistance and L the total inductance of the armature circuit. Note that R may include armature reaction and L mutual-inductance effects, as explained in Sec. 4.4. Combining Eqs. (4.36) and (4.37) gives then

$$\hat{i}_1 = \frac{k_g \hat{i}_f}{R - (k_g N_1 / N_f) + Ls} \quad (4.38)$$

The current i_1 is the field current for the main generator. If this generator is linear, then its open-circuit output voltage is

$$\hat{E}_o = k_2 \hat{i}_1 \quad (4.39)$$

where k_2 is the main-generator voltage constant. Also the control-field current i_f may be related to the control voltage E_f by Eq. (4.6). Combining Eqs. (4.38), (4.39), and (4.6), we obtain

$$\frac{\hat{E}_o}{\hat{E}_f} = \frac{(k_g / R_f) k_2}{(T_f s + 1)[R - (k_g N_1 / N_f) + Ls]} \quad (4.40)$$

The critical value for R is the value that makes

$$R = R_c = \frac{k_g N_1}{N_f} \quad (4.41)$$

Hence the transfer function for a critically tuned Rototrol becomes

$$\frac{\hat{E}_o}{\hat{E}_f} = \frac{k_g k_2}{R_f L} \frac{1}{s(T_f s + 1)} \quad (4.42)$$

The free s in the denominator substantiates the statement made above that the critically tuned Rototrol is an integrating device.

The effect of eddy currents in the iron of either the pilot generator or the main generator is fundamentally the same as for the simple generator.

For the pilot generator, Eq. (4.18) applies without change, and eddy currents simply increase the effective field time constant. For the main generator, it can be shown by arguments similar to those employed in connection with Eq. (4.18) that eddy currents simply decrease the total inductance L of the armature circuit. Demonstration is left to the reader.

Mutual inductance between the control circuit and the armature circuit also has essentially the same effect here as in the simple generator,

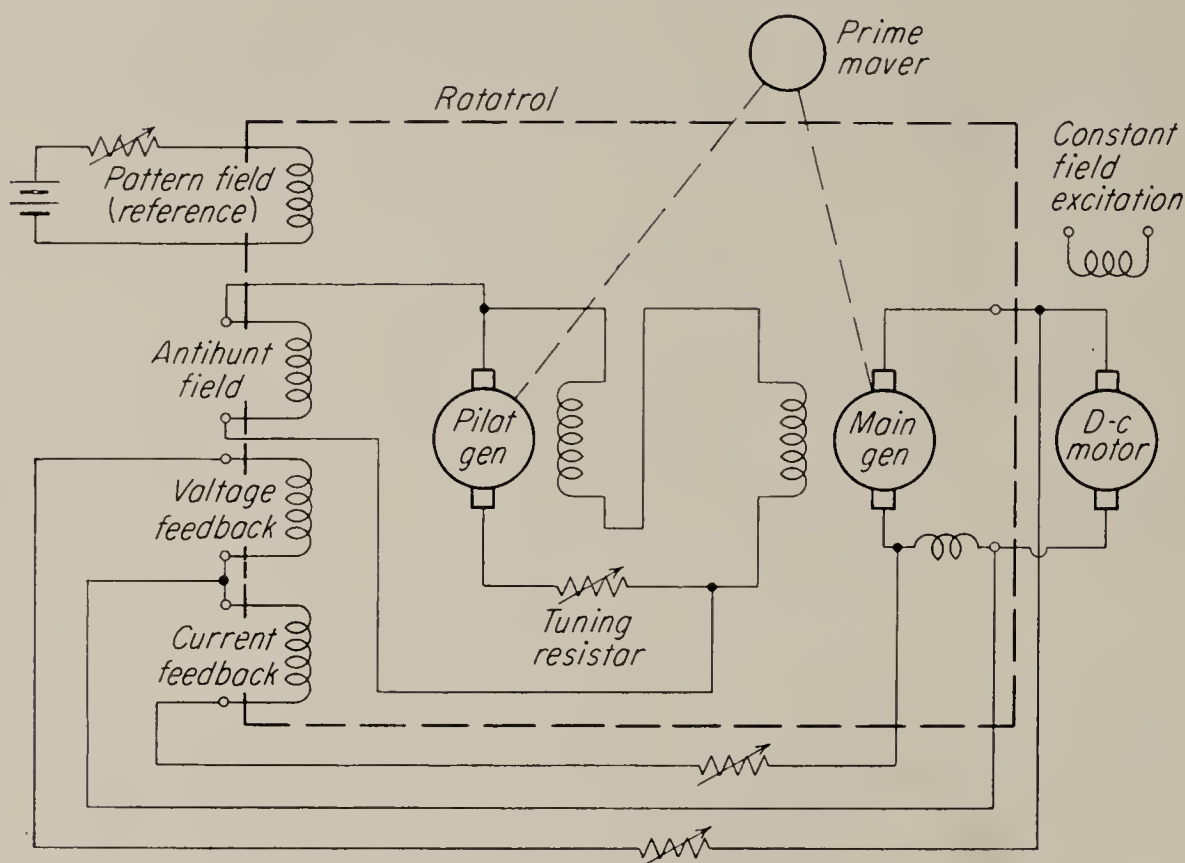


FIG. 4.13. A speed-control system using a Rototrol and d-c motor.

and an analysis similar to that carried out in connection with Eq. (4.28) leads to the result that Eq. (4.40) must be changed to read

$$\frac{\hat{E}_o}{\hat{E}_f} = \frac{(k_2/R_f)(k_g - M_{af}s)}{\left[(T_f s + 1)R - \frac{k_g N_1}{N_f} + Ls + \frac{(k_g - M_{af}s)M_{af}s}{R_f + L_f s} \right]} \quad (4.43)$$

when mutual inductance is taken into account. A check of polarities indicates that M_{af} in the above equation must be positive if the series winding is wound in a direction to aid the control field. Hence it appears that the Rototrol generator has a *non-minimum-phase* type of transfer function;¹ however, since the mutual-inductance effect is small, the non-minimum-phase root occurs at very high frequencies and is usually negligible. Equation (4.43) also indicates that mutual inductance does not affect the critical value of R , nor does it affect the integrating action of a critically tuned Rototrol.

¹ Bode, "Network Analysis and Feedback Amplifier Design," D. Van Nostrand Company, Inc., Princeton, N.J., 1945, pp. 242-244.

4.8. Pilot Generators with Several Control Windings. Pilot generators with more than one control winding are sometimes used in control systems where an electronic amplifier is undesirable. A typical system of this sort is shown in Fig. 4.13, where a Rototrol and d-c motor are used in a speed-regulating system.

To analyze the effect of a number of control fields on a generator, we consider for the moment a generator with two control fields, as shown in Fig. 4.14. The analysis is facilitated by the use of the superposition theorem, which permits us to replace first one and then the other of the two voltages by its internal impedance and to add the results of the individual computations. When one of the two voltages in Fig. 4.14 is replaced by its internal impedance, the resulting circuit is identical to Fig. 4.8. Hence the expression giving E_o as a function of the remaining input voltage will have the same form as Eq. (4.15). If the mutual coupling between the two fields is unity, we may then adapt Eq. (4.18) to obtain

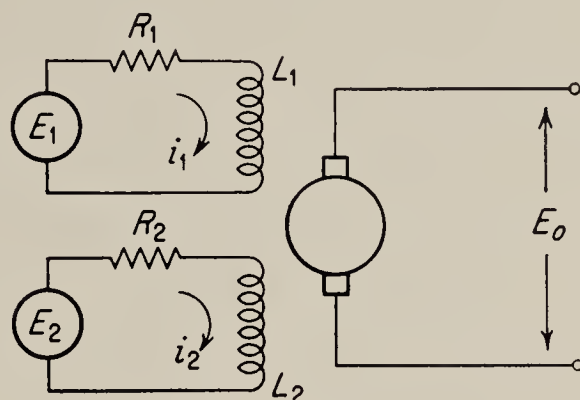


FIG. 4.14. Rototrol with two control fields.

$$\hat{E}_o = \frac{(k_1/R_1)\hat{E}_1 + (k_2/R_2)\hat{E}_2}{[(L_1/R_1) + (L_2/R_2)]s + 1} \quad (4.44)$$

Here k_1 and k_2 are voltage constants defined by

$$\begin{aligned} E_o &= k_1 i_1 && \text{if } i_2 = 0 \\ E_o &= k_2 i_2 && \text{if } i_1 = 0 \end{aligned}$$

and R_1 and R_2 represent the total resistance, including that of the voltage source in each of the two control circuits. This result may be generalized to more than two control circuits.

4.9. The Regulex Generator.

Instead of using a series field for the self excitation of the pilot generator, as in the Rototrol, it is also possible to use a shunt field. The resulting two-stage generator is called the *Regulex* and has been used in systems built by the Allis-Chalmers Company. A schematic

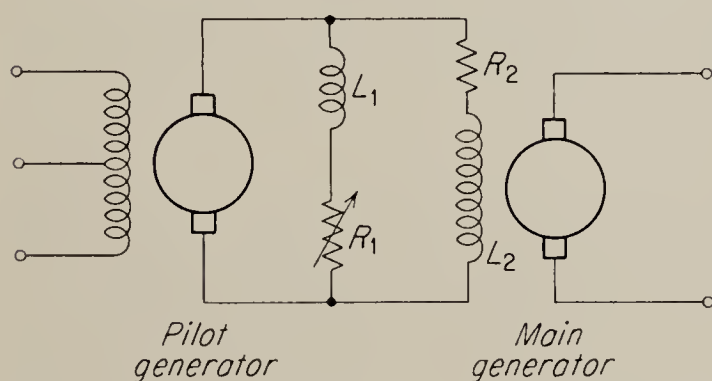


FIG. 4.15. A schematic of the Regulex generator.

diagram of this machine is shown in Fig. 4.15. Here again the pilot generator is shown with a push-pull field designed for excitation by an electronic amplifier; however this generator, like the Rototrol, is often built with several control fields. The theory of operation is

very similar to that of the Rototrol, and by proper adjustment of the shunt-field resistance R_1 , it is possible to obtain integrating action from this system. The complete transfer function has a slightly different form from that of the Rototrol; its computation is left to the reader.

4.10. The Amplidyne Generator. A generator often used in control systems is the Amplidyne, first described by Alexanderson, Edwards, and Bowman.¹ The *Amplidyne* is essentially a two-stage generator, like the

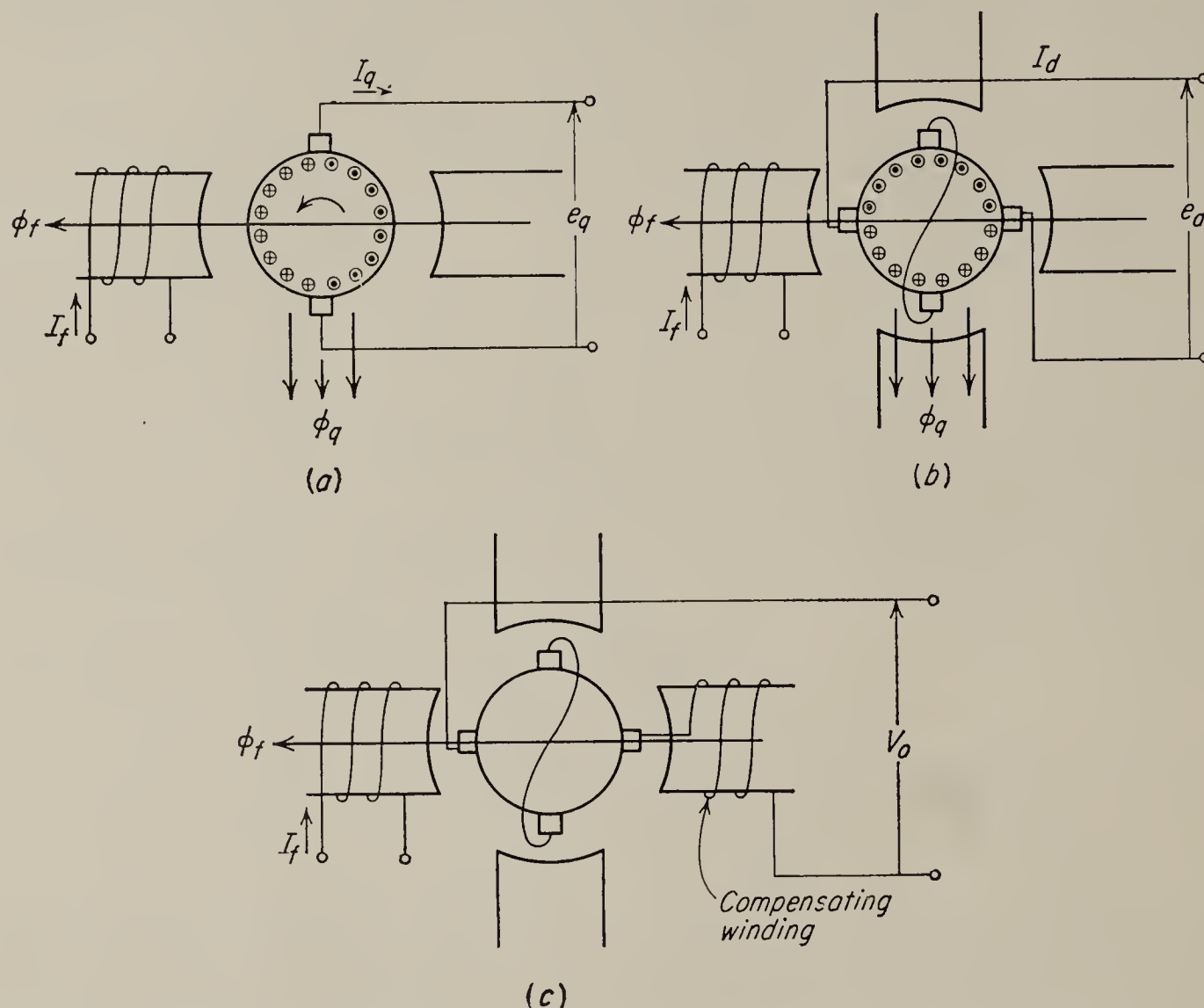


FIG. 4.16. Amplidyne principles: (a) ϕ_f is the flux set up by the control-field current I_f ; ϕ_q is caused by current I_q . (b) Quadrature brushes short-circuited; ϕ_d opposes ϕ_f , creating a constant-current generator. (c) Addition of compensating winding converts machine to Amplidyne.

Rototrol described in the previous sections; however the two stages are combined in a single machine with a single armature winding. The Amplidyne therefore tends to be somewhat smaller in size than a Rototrol of the same rating.

The principle of operation is as follows: Consider the simple d-c generator of Fig. 4.16a with a control-field current I_f . This current produces a flux ϕ_f , and if the machine rotates in the direction shown, the voltage e_q induced in the conductors will be in the direction indicated by the con-

¹ E. F. W. Alexanderson, M. A. Edwards, and K. K. Bowman, Dynamoelectric Amplifier for Power Control, *Trans. AIEE*, vol. 59, pp. 937-939, 1940.

ventional system of dots and crosses. If a current is taken from the brushes, an armature-reaction flux ϕ_q due to this current appears at right angles to ϕ_f . In conventional generators every attempt is made to suppress the armature reaction, but in the Amplidyne a magnetic path is provided to encourage it. Furthermore the brushes are short-circuited so that the maximum amount of ϕ_q is produced for a given control-field current. The armature-reaction flux will then also induce voltage in the armature conductors. This voltage is in a direction indicated by the dots and crosses of Fig. 4.16b. The second set of brushes shown in that figure is used to connect this voltage to the load. We refer to this voltage as the *direct-axis voltage* e_d . Since the quadrature flux ϕ_q is very much larger than the control flux ϕ_f , the direct-axis voltage is very much larger than the *quadrature-axis voltage* e_q . The additional amplification takes place in the magnetic circuit of the machine. The armature-reaction flux is somewhat analogous to the pilot-generator armature current of the Rototrol system.

When the output brushes of the machine shown in Fig. 4.16b are connected to a load, the resulting current flowing in the armature will again result in an armature-reaction effect, and the reader may convince himself that this effect opposes the control-field flux. The load current is therefore limited to a value such that the mmf of this armature reaction is equal to the magnetomotive force produced by the control field I_f . In other words the machine acts as a constant-current generator. Its load current is proportional to the control-field current and essentially independent of speed and direction of rotation of the armature. Referred to as a *Rosenberg generator*, this type of machine has been used for railroad-train lighting service since 1905, and as the *Metadyne* it has been used in diesel-electric locomotive drives.¹

In servo systems the constant-current feature is usually undesirable and is eliminated by passing the load current through a compensating field so located that the armature reaction of the load current is canceled by the mmf of the compensating field. This is shown in Fig. 4.16c. The addition of the compensating field converts the machine to an Amplidyne, a machine delivering a voltage that is only moderately affected by the load current. Amplidynes are occasionally built with a number of different control fields for purposes similar to those already discussed in connection with the Rototrol. Sometimes the quadrature brushes are not short-circuited directly; instead they are connected through field windings that may be located in either the quadrature or direct axis to obtain certain special operating characteristics such as, for instance, integrating action. A major problem in Amplidynes is sparking at the brushes, due to the armature reaction on which the operation of the machine depends.

¹ Pestarini, "Metadyne Statics," John Wiley & Sons, Inc., New York, 1952.

Special brushes and construction of the field poles are used to alleviate this difficulty. Nevertheless Amplidynes often spark very severely at the quadrature brushes when delivering large output voltages.

4.11. Analysis of the Amplidyne. The following analysis of the Amplidyne is based on a paper by Bower¹ and assumes complete linearity of the electric and magnetic circuits, neglects effects of eddy currents in the iron, and assumes the machine to be running at constant speed. The circuit under discussion is shown in Fig. 4.17, and the resistances and inductances indicated in that figure in all cases represent the entire resistance and inductance of each of the three circuits. We shall also be concerned with mutual effects between the circuits and shall distinguish

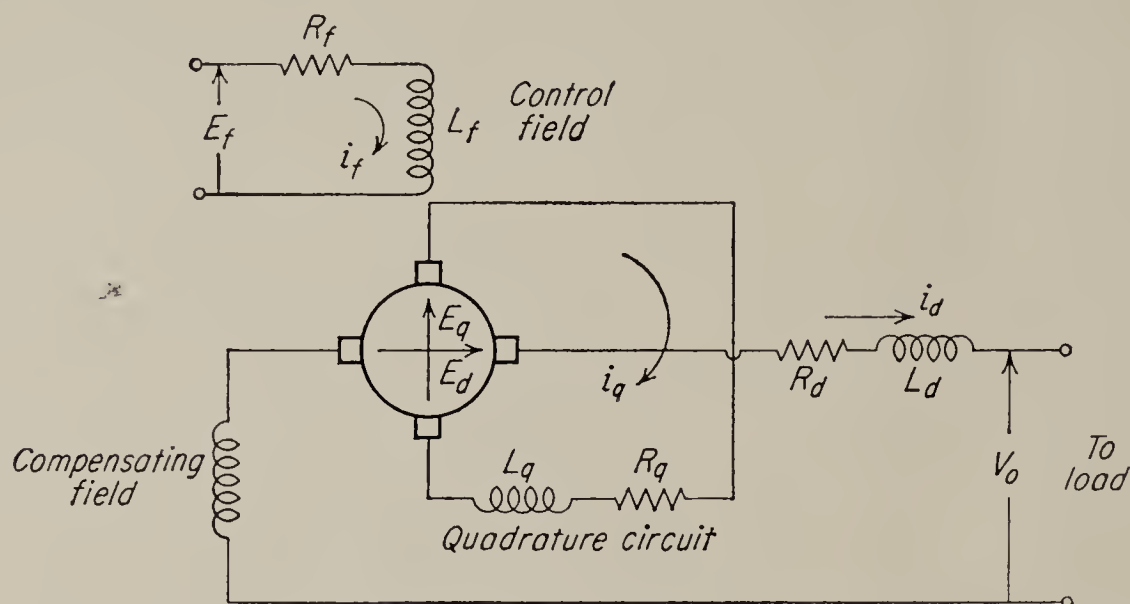


FIG. 4.17. Amplidyne circuit analyzed in text.

between two types. The first, designated by M , is the mutual inductance and relates voltage induced in one circuit to the rate of change of current in another circuit. The symbol N will be used to designate a speed-voltage parameter relating voltage generated in one circuit to the current in another circuit. To identify the various circuits we make use of a double-subscript notation using the subscripts f , q , and d to designate the control circuit, quadrature circuit, and direct circuit, respectively. Thus the symbol M_{fq} is the ratio of voltage in the quadrature circuit to the rate of change of control-field current causing it.

Before proceeding further with the analysis, we wish to show that, except for the mutual effects discussed above, the quadrature and direct circuits may be considered independent even though they use the same armature winding. The armature circuit has the form shown schematically in Fig. 4.18, where the brushes need not be exactly 90° apart. It is required only that each pair of brushes be exactly 180° apart. Under these conditions $Z_1 = Z_3$ and $Z_2 = Z_4$, and therefore $Z_1/Z_2 = Z_3/Z_4$.

¹ J. L. Bower, Fundamentals of the Amplidyne Generator, *Trans. AIEE*, vol. 64, pp. 873–81, 1945.

The circuit represents a balanced bridge, and therefore the voltage E_d is unaffected by i_q , and vice versa.

We now write the Kirchhoff mesh equation for each of the three circuits, using Laplace transform notation. For the control field we have

$$\hat{E}_f = (R_f + L_f s)\hat{i}_f - M_{qf}s\hat{i}_q - M_{df}s\hat{i}_d \quad (4.45)$$

In this equation M_{qf} is the mutual inductance between the control field and quadrature circuit and should be zero if the brushes are located exactly on the neutral axis and if the quadrature circuit does not have field windings collinear with the control circuit. In order to consider the possibility of shifted brushes, however, we retain this term and adopt the convention that a brush shift in the direction of rotation will make M_{qf} more positive. The situation is different for M_{df} , which is the mutual inductance between normally collinear fields and which is affected only slightly by brush shift. M_{df} is, however, a function of the effectiveness of the direct-axis compensating winding. Theoretically M_{df} would be zero if the compensation of the direct-axis armature reaction were perfect. In order to include the effect of imperfect compensation in the analysis, we adopt the convention that M_{df} is positive in an undercompensated machine, i.e., a machine for which the output voltage drops as load current is taken from it.

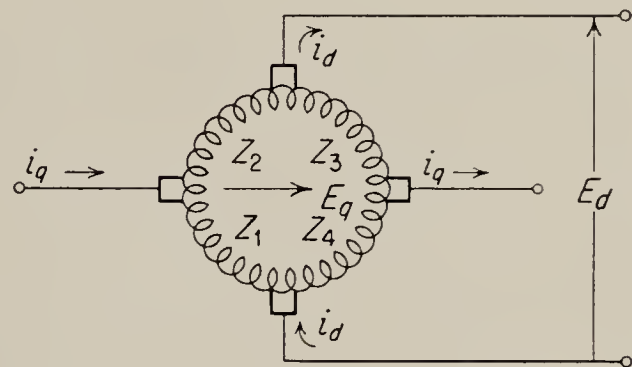


FIG. 4.18. Schematic of Amplidyne armature.

For the quadrature circuit we have

$$\hat{E}_q = N_{fq}\hat{i}_f = -M_{fq}s\hat{i}_f + (R_q + N_{qq} + L_q s)\hat{i}_q + (N_{dq} - M_{dq}s)\hat{i}_d \quad (4.46)$$

In this equation we assume (rather arbitrarily) that $E_q = N_{fq}\hat{i}_f$ is the “driving voltage” in the quadrature circuit which is opposed by the self- and mutual-impedance drops listed on the right side of the equation. Here again the mutual-inductance terms M_{fq} and M_{dq} are theoretically zero if the brushes are in the neutral axis and exactly 90° apart. M_{fq} must be the same as M_{qf} discussed above and becomes more positive as the brushes are shifted in the direction of rotation. However, since M_{dq} remains zero as long as the brushes are 90° apart, we adopt here the further convention that M_{dq} becomes positive if the output brushes are shifted relative to the quadrature-circuit brushes in the direction of rotation. The parameter N_{qq} represents speed voltage induced in the quadrature circuit by the quadrature current itself. This may be due to mmf produced in the direct axis by the quadrature current because of a series winding, or else it may simply be the effect of armature reaction in

the quadrature circuit. This parameter could be lumped with the resistance R_q to form an "effective" resistance R'_q , as discussed in connection with the simple generator (Sec. 4.4). N_{dq} represents the speed voltage in the quadrature circuit resulting from load current in the direct circuit. It is not much affected by brush shift but is a function of the amount of compensation supplied by the direct-circuit compensating coil. It is assumed to be positive for an undercompensated machine. Therefore, if N_{dq} is positive an increase of load current causes a decrease of the quadrature current and hence the output voltage.

For the direct circuit we have

$$\hat{E}_d = N_{qd}\hat{i}_q = (N_{fd} - M_{fd}s)\hat{i}_f - M_{qd}s\hat{i}_q + (R_d + N_{dd} + L_d s)\hat{i}_d + \hat{V}_o \quad (4.47)$$

This equation has the form of a generated voltage, $N_{qd}\hat{i}_q$, set equal to the sum of the impedance drop and the terminal voltage. All the remarks made previously concerning the signs of the mutual terms apply here without change. N_{dd} is the same type of parameter as N_{qq} in Eq. (4.46). $N_{fd}\hat{i}_f$ is the speed voltage generated in the direct circuit by direct coupling from the control field. It is zero if the brushes are on the neutral axis. N_{fd} is positive when the brushes are shifted in the direction of rotation.

The primary purpose of this analysis is the determination of the Thévenin equivalent of an Amplidyne; hence we wish to solve Eqs. (4.45) to (4.47) in such a way that the open-circuit voltage and output impedance become evident. This is most easily done by rearranging these equations in the form

$$\begin{aligned} \hat{E}_f - \hat{i}_d Z_{df} &= Z_{ff}\hat{i}_f + Z_{qf}\hat{i}_q \\ -\hat{i}_d Z_{dq} &= Z_{fq}\hat{i}_f + Z_{qq}\hat{i}_q \\ -\hat{i}_d Z_{dd} &= Z_{fd}\hat{i}_f + Z_{qd}\hat{i}_q + \hat{V}_o \end{aligned} \quad (4.48)$$

where, by comparison with Eqs. (4.45) to (4.47), we have

$$\begin{aligned} Z_{ff} &= R_f + L_f s \\ Z_{fq} &= -N_{fq} - M_{fq}s \\ Z_{fd} &= N_{fd} - M_{fd}s \\ Z_{qf} &= -M_{qf}s \\ Z_{qq} &= R_q + N_{qq} + L_q s \\ Z_{qd} &= -N_{qd} - M_{qd}s \\ Z_{df} &= -M_{df}s \\ Z_{dq} &= N_{dq} - M_{dq}s \\ Z_{dd} &= R_d + N_{dd} + L_d s \end{aligned} \quad (4.49)$$

The solution of Eq. (4.48) is shown below in determinant form:

$$\hat{V}_o = \begin{vmatrix} Z_{ff} & Z_{qf} & 1 \\ Z_{fq} & Z_{qq} & 0 \\ Z_{fd} & Z_{qd} & 0 \end{vmatrix} \hat{E}_f - \begin{vmatrix} Z_{ff} & Z_{qf} & Z_{df} \\ Z_{fq} & Z_{qq} & Z_{dq} \\ Z_{fd} & Z_{qd} & Z_{dd} \end{vmatrix} \hat{i}_d \quad (4.50)$$

and is of the form

$$\hat{V}_o = H\hat{E}_f - Z_o\hat{i}_d$$

where H is the no-load transfer function and Z_o is the output impedance of the Amplidyne. $H\hat{E}_f$ may be considered the Thévenin open-circuit voltage.

Although Eq. (4.50) is a complete description of Amplidyne performance, subject to the assumptions made at the beginning of this paragraph, it is too cumbersome to be of much practical use. It is therefore necessary to make a number of simplifications. These simplifications are made on the following bases: (1) In accordance with the discussion on mutual inductance in Sec. 4.4, where it was shown that mutual-inductance effects are often negligible compared to speed effects in generators, we may neglect a number of mutual-inductance terms in Eq. (4.50) where this is desirable. (2) A number of coupling terms in Eq. (4.50) are zero when the brushes are located properly in the neutral axis and can, therefore, be considered negligibly small if only small shifts are permitted.

The expression for the open-circuit voltage becomes much more manageable if we let $Z_{qf} = -M_{qf}s = 0$; this simplification is justified on both the grounds given above. Then the open-circuit voltage becomes

$$\hat{E}_o = \left[\frac{N_{fq}N_{qd}/R_f(R_q + N_{qq})}{(T_fs + 1)(T_qs + 1)} - \frac{N_{fd} - M_{fd}s}{R_f(1 + T_fs)} \right] \hat{E}_f \quad (4.51)$$

where $T_f = \frac{L_f}{R_f}$, the control-circuit time constant, and $T_q = \frac{L_q}{(R_q + N_{qq})}$, the quadrature-circuit time constant. In this expression the term of primary importance, and the only one usually considered, is the first. It indicates that the output voltage is a function of the square of the speed (since both N 's are proportional to speed) and that the open-circuit transfer function is characterized by two time constants. Brush shift has an effect on this term through the parameter N_{qq} , which increases with brush shift in the direction of rotation and which may become negative for brush shifts in the opposite direction. N_{qq} can also be made negative by a series coil in the quadrature circuit so arranged that its mmf is in a direction to aid the mmf of the control field. In this way N_{qq} can be

made exactly equal and opposite to R_q , and the open-circuit output voltage becomes approximately

$$\hat{E}_o = \hat{E}_f \frac{N_{fq}N_{qd}/R_fL_q}{s(T_fs + 1)} \quad (4.52)$$

indicating that the output voltage (at low frequencies) is the integral of the input voltage. If an Amplidyne is to be used in this way, a series coil with an adjustable shunting resistor may be used to adjust N_{qq} .

The second term of Eq. (4.51) represents the component of the output voltage resulting from direct coupling between the control field and the output circuit. If the brushes are not shifted from their neutral position, N_{fd} is theoretically zero; hence the only coupling remaining is that caused by mutual inductance. This will also be quite small in an Amplidyne that is properly compensated. Thus, the entire term is usually neglected. It should be pointed out, however, that while the resulting *low-frequency approximation* to the Amplidyne transfer function is usually satisfactory, one occasionally finds an unusual system in which there is a minor loop or a parasitic loop with a wide passband. In such a system the presence of a term like the second term of Eq. (4.51) that does not vanish at high frequencies may make the difference between a stable system and one that oscillates at high frequencies.

The output impedance is given by a much more complicated expression than the open-circuit voltage, and our simplifications must therefore become more extensive. We assume that the brushes are located exactly on the neutral axis, so that

$$\begin{aligned} Z_{fq} &= -N_{fq} & Z_{qd} &= -N_{qd} \\ Z_{fd} &= -M_{fd}s & Z_{dq} &= N_{dq} \\ Z_{qf} &= 0 \end{aligned} \quad (4.53)$$

with the other Z 's as in Eq. (4.49). With these simplifications

$$\begin{aligned} Z_o = R_d + N_{dd} + L_ds + \frac{N_{dq}N_{qd}}{R_q + N_{qq}} - \frac{N_{df}N_{fq}N_{qd}}{R_f(R_q + N_{qq})} s \\ - \frac{M_{df}^2s^2}{R_f} \quad (4.54) \end{aligned}$$

In this expression the first three terms represent simply the self-impedance of the load circuit, and it is almost obvious that it would be a component of the output impedance. The first fractional term represents the effect of imperfect compensation on the output impedance. It is positive in an undercompensated Amplidyne, and in many commercial machines it

is several times larger than the first term at low frequencies. This term may be represented as the impedance of a resistance in parallel with a capacitance, as shown in Fig. 4.19, and it may be so represented in an

$$R = \frac{N_{dq}N_{qd}}{R_q + N_{qq}}$$

$$RC = T_q$$

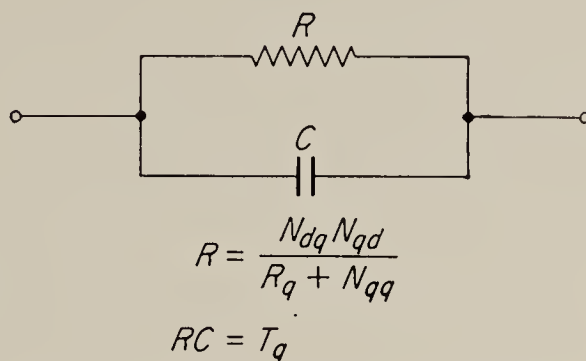


FIG. 4.19. Equivalent circuit of output impedance of Amplidyne.

equivalent circuit of the Amplidyne. The last two terms are usually omitted because they are primarily due to mutual inductance and therefore are negligible compared to the other two terms. However, here again these terms must sometimes be retained in the “unusual” system.

4.12. The Equivalent Circuit of the Amplidyne. If we consider only the first term of Eq. (4.51) for the open-circuit voltage and only the first four terms of Eq. (4.54) for the output impedance, the equivalent circuit shown in Fig. 4.20 results. Typical values for the time constants are:

$T_f = \frac{L_f}{R_f}$, between 0.1 and 0.2 sec. May however be reduced to a few milliseconds by use of high-resistance drive, such as pentode vacuum tube

$T_q = \frac{L_q}{R_q + N_{qq}}$, between 0.02 and 0.07 sec

$T_d = \frac{L_d}{R_d + N_{dd}}$, between 0.005 and 0.05 sec

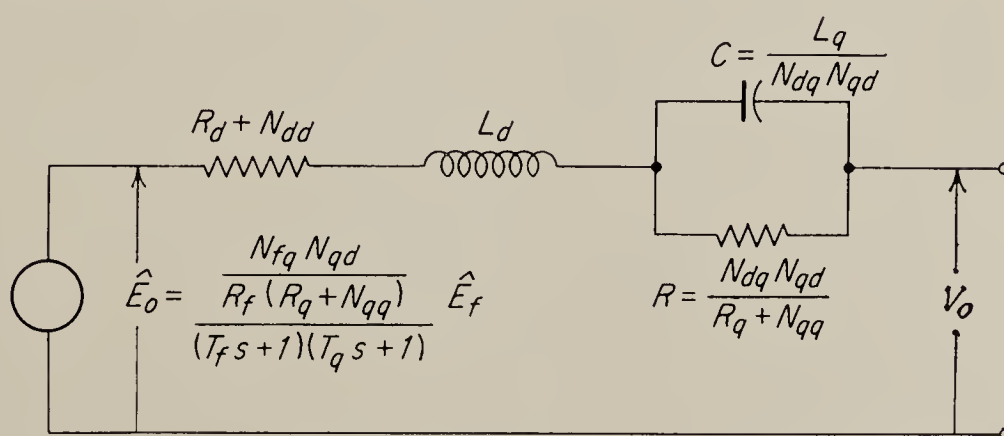


FIG. 4.20. Approximate equivalent circuit of the Amplidyne.

The values of gain and output resistance depend, of course, on the size and other construction features.

The reader should note that one effect of the parallel RC circuit in the equivalent circuit is that the quadrature-circuit time constant T_q does not necessarily appear in the transfer function of a loaded machine. The demonstration of this fact is left as an exercise (see Probs. 4.3 and 4.4).

4.13 Inaccuracies in the Amplidyne Analysis. One of the assumptions made at the beginning of the analysis was that both the electric and magnetic circuits of the Amplidyne were linear. Without this assumption the analysis could not have been carried through. However, such an assumption is at considerable variance with observed facts. In the Amplidyne both magnetic hysteresis and the nonlinearities caused by the brush contact serve to alter the performance from that expected by a linear analysis. We wish to consider particularly the latter effect, since it is somewhat unusual.

In most generators it is perfectly proper to neglect the nonlinear brush resistance in comparison with the usually very much larger resistance encountered in the load circuit. In the Amplidyne quadrature circuit,

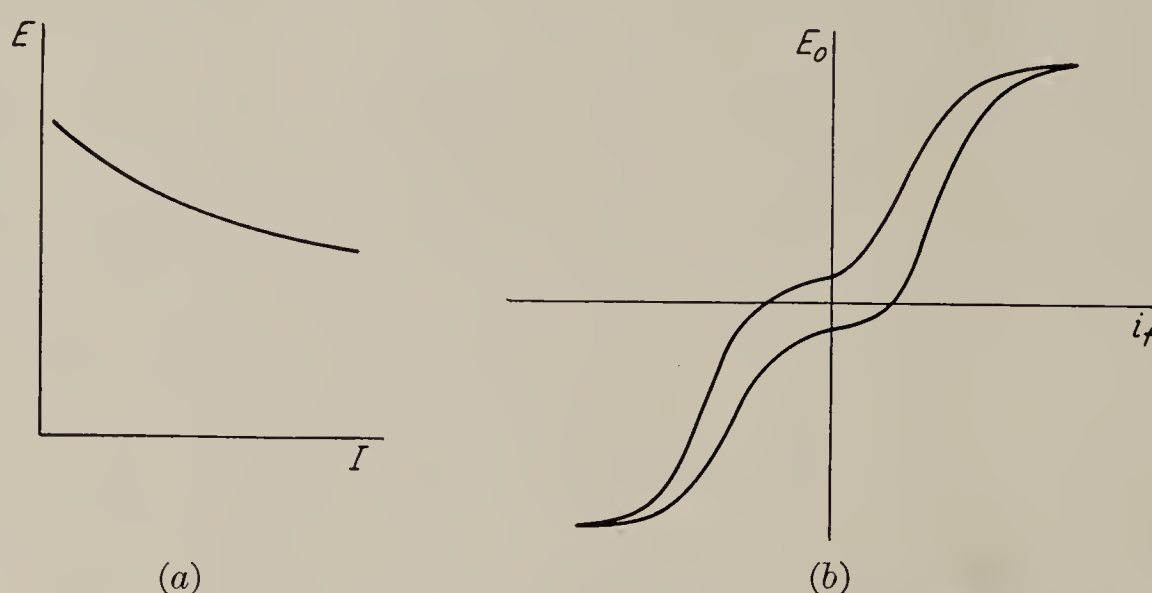


FIG. 4.21(a) Brush voltage-current characteristic; (b) actual Amplidyne hysteresis curve.

however, the brush-contact drop is a very large part of the total voltage drop, since the brushes are short-circuited. The exact nature of the brush-contact voltage depends on the brush material and contact pressure. The general shape of the voltage as a function of current is similar, however, to that found in electric arcs: a high voltage at small currents, decreasing as the current increases (see Fig. 4.21a). The result is that a small quadrature voltage can produce only a very small quadrature current, but as the quadrature voltage increases, the current becomes larger much more rapidly, sometimes showing a sort of trigger action due to the brush-voltage function. As a result of this action, the over-all hysteresis curve (output voltage versus control-field current) characteristically has a shape similar to that shown in Fig. 4.21b, with a deadband effect at the center. When an Amplidyne is used in a control system, this deadband tends to reduce the loop gain and the accuracy as the input and output approach correspondence. There may also be a tendency to hunt, if the deadband is very severe.

4.14. Comparison of the Various Generator Types. A question of considerable practical interest is which of the three generators, the Amplidyne, the Rototrol, or the Regulex, is "best" for the application at hand. While we can give no complete answer to this question, a short discussion relating the various features of the machines may be helpful.

The Rototrol and Regulex are very similar. They are essentially two simple generators operated in a cascade connection, equipped with a special feedback connection that makes the output proportional to the integral of the input. In other respects their characteristics do not differ from those of two simple generators designed to be used together. They would have the same maximum power output, the same overload capacity, the same size and weight, and about the same speed of response. The integrating feature is of interest primarily in control systems that do not use electronic amplifiers to drive the generator fields. If electronic control amplifiers are used, the integration can be performed more easily and probably more accurately within the control amplifier. Thus, if a given power output is required, there seems little reason to choose either the Rototrol or the Regulex in preference to two simple generators.

The Amplidyne is also essentially a two-stage generator; however, the two stages are combined on a single armature. In performance the Amplidyne is probably somewhat poorer than two separate generators. This is due to the various interactions taking place inside the machine. The problem of rather heavy sparking at the brushes has already been mentioned. Furthermore, although a number of different characteristics may be obtained from the Amplidyne by brush shifting or different winding arrangements, it seems clear that two separate generators driven by an electronic amplifier should permit much greater flexibility in the adjustment of the characteristics to meet a particular requirement. The chief advantage of the Amplidyne is undoubtedly its compactness. Since it uses only a single yoke and armature, it occupies less space and weighs less than two equivalent cascaded generators would. Hence Amplidynes are often preferred in aircraft applications.

4.15. D-C Motors with Separate Excitation. When one of the d-c generators discussed in the previous section is used to control the speed of a d-c motor, the arrangement used is usually the familiar Ward-Leonard system shown in Fig. 4.1, where the motor field is separately excited by a fixed supply and the two armatures are connected. The transfer function relating the speed or position of the motor shaft to the control signal applied to the generator will depend on the type of generator used. Since all the generators discussed thus far may be replaced by equivalent circuits of the Thévenin type, it is only necessary to find the general transfer function of a motor connected to such an equivalent generator.

The linear analysis of motors requires (1) the assumption of viscous friction at the output shaft, (2) linearity of the electric and magnetic circuits, (3) negligibly small armature reaction. For the moment, no attempt will be made to justify these assumptions, but the effects of the nonlinearities and armature reaction will be considered later.

The system shown in Fig. 4.22 will be considered. The motor is connected to a generator with open-circuit voltage E_o and output impedance Z_o . It has an armature resistance R_m and an inductance L_m . The armature has a moment of inertia J , and the coefficient of viscous friction between the armature shaft and a stationary reference system is B . The armature current is i_a , and the motor shaft position is θ . The motor is connected to a load requiring a load torque Q_L .

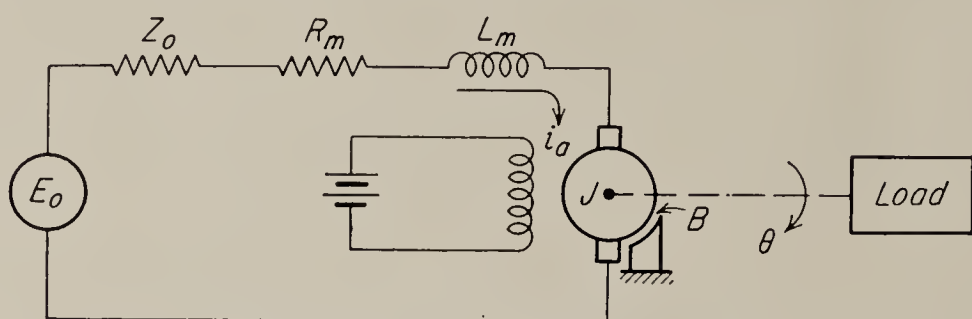


FIG. 4.22. D-c motor with armature control.

The counter emf is given by the same equation as the open-circuit generated voltage of a generator, Eq. (4.1); however, here N_a , P , ϕ , and p are all constant, and $\Omega = \dot{\theta}$ is the independent variable. Therefore this equation is written in the form

$$\hat{E}_c = k_v \hat{\theta} = k_v s \hat{\theta} \quad (4.55)$$

The torque developed by a motor is given in electric machinery texts¹ as $Q_d = k\phi i_a$, where k is a constant containing the same sort of parameters as are contained in Eq. (4.1). Again, if the magnetic flux ϕ is constant, this can be written as

$$\hat{Q}_d = k_t \hat{i}_a \quad (4.56)$$

A fixed relationship exists between k_t and k_v , as is indicated by the following argument: In electrical terms the power developed is the product of counter emf and armature current. In mechanical terms it is the product of developed torque and speed. These two expressions must be equal, but they are usually expressed in different units. Hence we have

$$E_c i_a = C Q_d \dot{\theta} \quad (4.57)$$

where C is the factor converting electrical to mechanical units of power. Using Eqs. (4.55) and (4.56), this gives

$$k_v = C k_t \quad (4.58)$$

¹ For instance, Dawes, "A Course in Electrical Engineering," vol. I, "Direct Currents," 4th ed., McGraw-Hill Book Company, Inc., New York, 1952, p. 485.

The value of C is unity if a consistent set of units such as the mks system is used.

The armature current is given by

$$\hat{i}_a = \frac{\hat{E}_o - \hat{E}_c}{Z_o + R_m + L_ms} \quad (4.59)$$

Finally the sum of all torques applied to the motor shaft must be zero, or

$$\hat{Q}_d = k_t \hat{i}_a = Js^2 \hat{\theta} + Bs \hat{\theta} + \hat{Q}_L \quad (4.60)$$

The three equations, (4.55), (4.59), and (4.60), contain five unknowns; hence we can solve for three as a function of the other two. We consider \hat{E}_o and \hat{Q}_L the independent variables and solve for $\hat{\theta}$. The solution is

$$\hat{\theta} = \frac{(\hat{E}_o/k_v) - (\hat{Q}_L Z/k_v k_t)}{s[(JZ/k_v k_t)s + (BZ/k_v k_t) + 1]} \quad (4.61)$$

where Z is the total armature-circuit impedance: $Z = Z_o + R_m + L_ms$. Note the free s in the denominator of this expression; its presence indicates that a motor acts as an integrating device for the conditions given. This is simply due to the fact that the shaft position is the integral of shaft speed.

It is worth pointing out that Eq. (4.61) has the form

$$\hat{\theta} = H\hat{E}_o - Z_m \hat{Q}_L \quad (4.62)$$

in which $H\hat{E}_o$ is the no-load speed and Z_m might be referred to as a *mechanical output impedance*. This means that a motor may be represented by a *mechanical Thévenin equivalent*, a concept that is occasionally useful.

In many cases of practical importance the armature-circuit impedance may be considered to be resistive. This is a very good assumption when a simple generator or a Rototrol is used to drive the motor, and even with an Amplidyne drive the results are usually not far in error. If this simplification is made and if the total resistance of the armature circuit is R , Eq. (4.61) indicates that for $E_o = 0$ and Q_L constant, the motor speed will reach the value

$$\dot{\theta} = \frac{Q_L}{B + (k_v k_t/R)} = \frac{Q_L}{(k_v k_t/R)[1 + (BR/k_v k_t)]} \quad (4.63)$$

The denominator term $B + k_v k_t/R$ represents the effective coefficient of viscous damping of the motor. It is made up of the coefficient of mechanical friction, B , and the term $k_v k_t/R$, which represents the electrical damping. In most motors, the electrical damping is very much larger than the mechanical friction, typical values of the ratio $BR/k_v k_t$ being of the order of 0.01. Hence the term $BZ/k_v k_t$ in Eq. (4.61) may be omitted, and the

transfer function takes the approximate form

$$\hat{\theta} = \frac{\hat{E}_o/k_v}{s[(JR/k_vk_t)s + 1]} - \frac{\hat{Q}_LR/k_vk_t}{s[(JR/k_vk_t)s + 1]} \quad (4.64)$$

In this equation the factor JR/k_vk_t is a time constant referred to as the *motor-inertia time constant*, and it has a value ranging from about 0.01 to 0.1 sec in motors of standard design connected to a zero-impedance source.

4.16. Relation between the Transfer Function and the Speed-Torque Curves. The equation for the speed-torque curves of a d-c motor may be obtained from the complete motor transfer function [Eq. (4.61)]. In order to do this, first multiply Eq. (4.61) through by s to convert the output to speed rather than position and then set s equal to zero. The last step is required because speed-torque curves are static characteristics, and it changes the Z in Eq. (4.61) into R_a , the total armature-loop resistance. The equation for the speed-torque curves becomes

$$\dot{\theta} = k_1E_o - k_2Q_L \quad (4.65)$$

where

$$k_1 = \frac{1}{k_v(1 + BR_a/k_vk_t)} \quad (4.66)$$

and

$$k_2 = \frac{R_a}{BR_a + k_vk_t} \quad (4.67)$$

Since all the parameters defining k_1 and k_2 are constant in a linear motor, the speed-torque curves are parallel straight lines, and they are equidistant for equal increments of E_o . A typical set of such speed-torque curves is shown in Fig. 4.23.

Of greater practical interest than the determination of the speed-torque curves from the transfer function is the reverse process, namely, obtaining the transfer function from the speed-torque curves. This is often a convenient procedure in practice when the speed-torque curves have been measured or are otherwise available. The moment of inertia of the motor armature must be known. Suppose that it has the value J and suppose an external torque Q_o to be applied to the motor. Setting Q_L in Eq. (4.65) equal to $Q_o + J d\dot{\theta}/dt$ and using Laplace transform notation, we obtain

$$s\hat{\theta} = k_1\hat{E}_o - k_2Js^2\hat{\theta} - k_2\hat{Q}_o$$

or

$$\hat{\theta} = \frac{k_1\hat{E}_o - k_2\hat{Q}_o}{s(k_2Js + 1)} \quad (4.68)$$

If the values of k_1 and k_2 from Eqs. (4.66) and (4.67) are substituted into Eq. (4.68), it will be found that an expression almost identical in form with Eq. (4.61) is obtained. The only difference is that the armature impedance Z in Eq. (4.61) has been replaced by the resistance R_a . Thus the transfer function that has been obtained from the speed-torque curves is correct if the armature inductance is negligible.

The procedure outlined above may be used to determine the approximate transfer function of any power element for which the speed-torque curves are known. It is particularly useful in obtaining the quasi-linear transfer function of nonlinear power units. In defining this transfer function, the assumption is made that the characteristics of the unit are sufficiently regular that, for small variations of the input and output variables, they may be replaced by straight lines. The quasi-linear transfer function is usually a function of the operating point about which the small variations take place.

To illustrate the method of obtaining the transfer function, consider the speed-torque curves of a typical nonlinear device, say, a d-c motor, driven

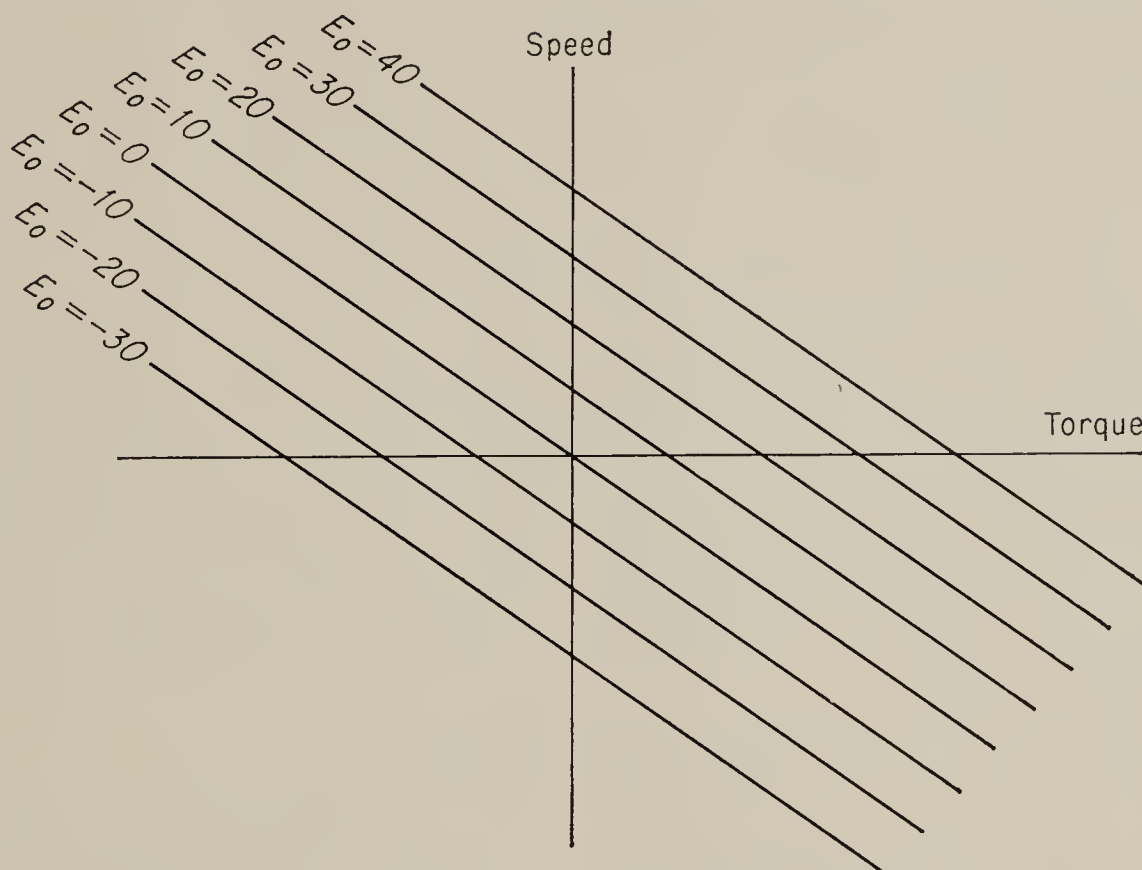


FIG. 4.23. Speed-torque curves of linear d-c motor.

by a nonlinear source of some sort (see Fig. 4.24). The curves are assumed to be plotted for integral increments of some input variable x corresponding to the generated voltage E_o in Eq. (4.65). At the operating point P the average vertical distance between the curves is equal to the parameter k_1 , while the slope of the particular curve passing through P is $-k_2$. By direct analogy with Eq. (4.68) the transfer function at the operating point P is then given by

$$\hat{\theta} = \frac{k_1 \hat{x} - k_2 \hat{Q}_o}{s(k_2 J s + 1)} \quad (4.69)$$

Since the speed-torque curves are not linear and not equidistant, the parameters k_1 and k_2 will vary with the operating point.

It should be clear from the above discussion that the transfer function obtained from the speed-torque curves can show only a single time con-

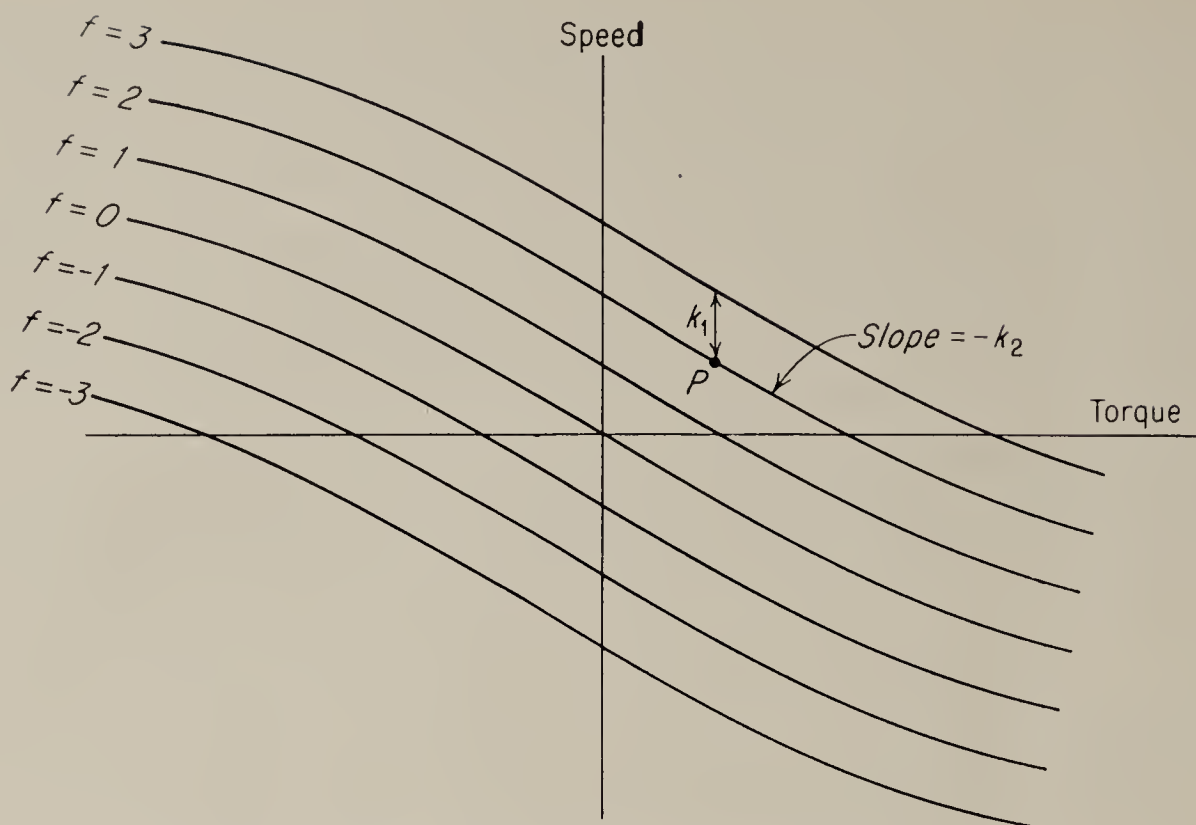


FIG. 4.24. Nonlinear speed-torque curves for d-c motor.

stant, namely, the inertia time constant. Its accuracy depends, therefore, on the assumption that other time constants are negligible.

4.17. An Equivalent Circuit for the D-C Motor. It is sometimes convenient to replace the motor by an equivalent electric circuit element. Such an element can be found if Eqs. (4.55), (4.59), and (4.60) are solved for the armature current in terms of E_o and Q_L . The result may be put in the form

$$\hat{i}_a = \frac{\hat{E}_o + \frac{\hat{Q}_L/k_t}{(J/k_vk_t)s + B/k_vk_t}}{Z_o + R_m + L_ms + \frac{1}{(J/k_vk_t)s + (B/k_vk_t)}} \quad (4.70)$$

If for the moment we ignore the load torque, it is seen that the denominator represents an impedance consisting of the elements Z_o , R_m , L_ms , all

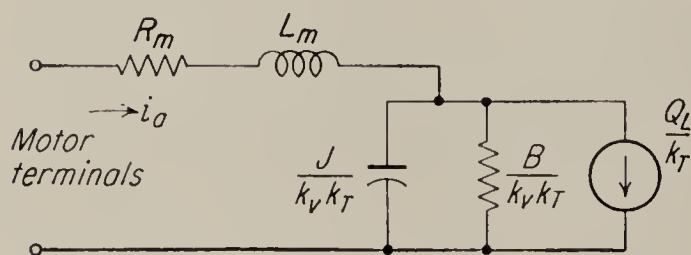


FIG. 4.25. An equivalent circuit of a d-c motor.

connected in series with a capacitor J/k_vk_t in parallel with a resistance B/k_vk_t . Furthermore, it may be demonstrated that a current source \hat{Q}_L/k_t connected in parallel with the combination J/k_vk_t and B/k_vk_t properly represents the load. The circuit therefore takes the form of

Fig. 4.25, where we have omitted the output impedance Z_o of the generator. It can be shown that the voltage across the capacitor is the counter emf of the motor, and the speed can therefore be obtained by dividing the capacitor voltage by k_v .

If friction is neglected and if there is no shaft load, the equivalent circuit is a simple RLC series circuit, and all equations pertaining to the motor will have the form of the equations of this circuit.

4.18. Inaccuracies in the Motor Analysis. The assumptions of linearity, etc., that were made to facilitate the analysis will now be examined briefly. Nonlinearity of the electric circuit is confined primarily to the brush-contact resistance and will cause a slight deadband effect in the motor's curve of speed versus input voltage. This effect was discussed in connection with the Amplidyne (Sec. 4.13) but is so small in a motor that it is usually neglected. A much more pronounced deadband effect is produced by the static friction between the rotating and stationary parts of the motor; it is sometimes serious enough, particularly in small motors, to cause instability of the servo of which the motor is a part. Aside from this effect, however, the fact that the friction drag is rarely proportional to speed is of no great consequence, since it was shown that the friction term can usually be neglected in the motor transfer function.

Armature reaction in motors results in a reduction of flux, just as it does in generators; hence both k_t and k_v are reduced slightly when large armature currents flow. It is seen therefore that armature reaction in a motor cannot simply be absorbed in an equivalent resistance as it is in the generator. However, since the field is usually operated in saturation and since compensating windings and commutating poles are often used on motors to improve commutation, armature reaction is not usually a major factor in motor performance.

Nonlinearity of the magnetic circuit can have no effect on the performance, since the motor field is maintained (in the absence of armature reaction) at a constant strength.

4.19. Other D-C Motor Types. A motor that is fairly often used in low-power servo applications is the split-field series motor, a circuit of which is shown in Fig. 4.26. These motors are usually driven directly from a push-pull electronic amplifier (see Sec. 3.13), or by a polarized relay, as shown in Fig. 4.28, rather than from a generator. For this reason their power output is limited to about 50 watts. The two series field coils are wound in such a way that, for the current directions shown in Fig. 4.26, the mmf's oppose; hence the flux is approximately proportional to the current difference, if the magnetic circuit is assumed to be linear. Since the armature current is the sum of the two field currents, the developed torque is

$$Q_d = k_t(I_1 + I_2)(I_1 - I_2) \quad (4.71)$$

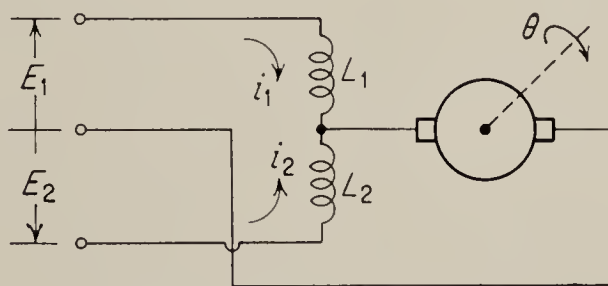


FIG. 4.26. The split-field series motor.

and the counter emf may be written as

$$E_c = k_v \dot{\theta}(I_1 - I_2) \quad (4.72)$$

The two currents I_1 and I_2 may be found by application of Kirchhoff's laws. In order to simplify the expressions, let us define a self-impedance Z_s as the impedance measured between one of the outside terminals and the center terminal with the motor not running and the remaining terminal open-circuited. This self-impedance should be the same for both windings and is of the form $R + j\omega L$. R and L are the sums of the resistances and inductances, respectively, of the armature and one of the field windings. Let us also define the mutual impedance

$$Z_M = \left. \frac{\partial E_1}{\partial I_2} \right|_{I_1, \dot{\theta}=0} = \left. \frac{\partial E_2}{\partial I_1} \right|_{I_2, \dot{\theta}=0} \quad (4.73)$$

The mutual impedance will be of the form $R_a + j\omega L_a + j\omega M$, where R_a and L_a are armature resistance and inductance, respectively, and M is the mutual inductance between the two windings. With these definitions and the use of Eq. (4.72), we have

$$\begin{aligned} E_1 &= Z_s I_1 + Z_M I_2 + k_v \dot{\theta}(I_1 - I_2) \\ E_2 &= Z_M I_1 + Z_s I_2 + k_v \dot{\theta}(I_1 - I_2) \end{aligned} \quad (4.74)$$

Note that these equations *are not* written in Laplace transform notation, since they contain products of the variables. Solving (4.74) for I_1 and I_2 , we get

$$\begin{aligned} I_1 &= \frac{E_1 Z_s - E_2 Z_M - k_v \dot{\theta}(E_1 - E_2)}{Z_s^2 - Z_M^2} \\ I_2 &= \frac{E_2 Z_s - E_1 Z_M - k_v \dot{\theta}(E_1 - E_2)}{Z_s^2 - Z_M^2} \end{aligned} \quad (4.75)$$

Let us suppose now that E_1 and E_2 vary an equal and opposite amount from a quiescent value E_0 ; i.e.,

$$\begin{aligned} E_1 &= E_0 + \Delta E \\ E_2 &= E_0 - \Delta E \end{aligned} \quad (4.76)$$

Then the substitution of (4.76) and (4.75) in (4.71) gives for the developed torque

$$Q_d = \frac{4k_t E_0 \Delta E}{(Z_s^2 - Z_M^2)} - \frac{8k_t k_v \dot{\theta} (\Delta E)^2}{(Z_s^2 - Z_M^2)(Z_s - Z_M)} \quad (4.77)$$

It may be observed that, at standstill, the developed torque is directly proportional to ΔE and that, for any fixed value of ΔE , torque and speed are related by a straight-line function. Speed-torque curves for various values of ΔE are shown in Fig. 4.27.

Since the motor is seen to be quite nonlinear, it is impossible to find an exact linear transfer function for it. If, however, operation is confined to low speeds, as may often be the case in servo systems, and if $E_0 \gg \Delta E$, an approximate transfer function can be derived by neglecting the second term of Eq. (4.77) and equating the developed torque to the sum of all other torques acting on the motor; i.e.,

$$\hat{Q}_d = \hat{Q}_L + Js^2\hat{\theta} + Bs\hat{\theta} \quad (4.78)$$

For operation at low frequencies, $Z_s + Z_M \approx R_f + 2R_a$ and $Z_s - Z_M \approx R_f$, where R_f and R_a are the resistances of one field coil and the armature,

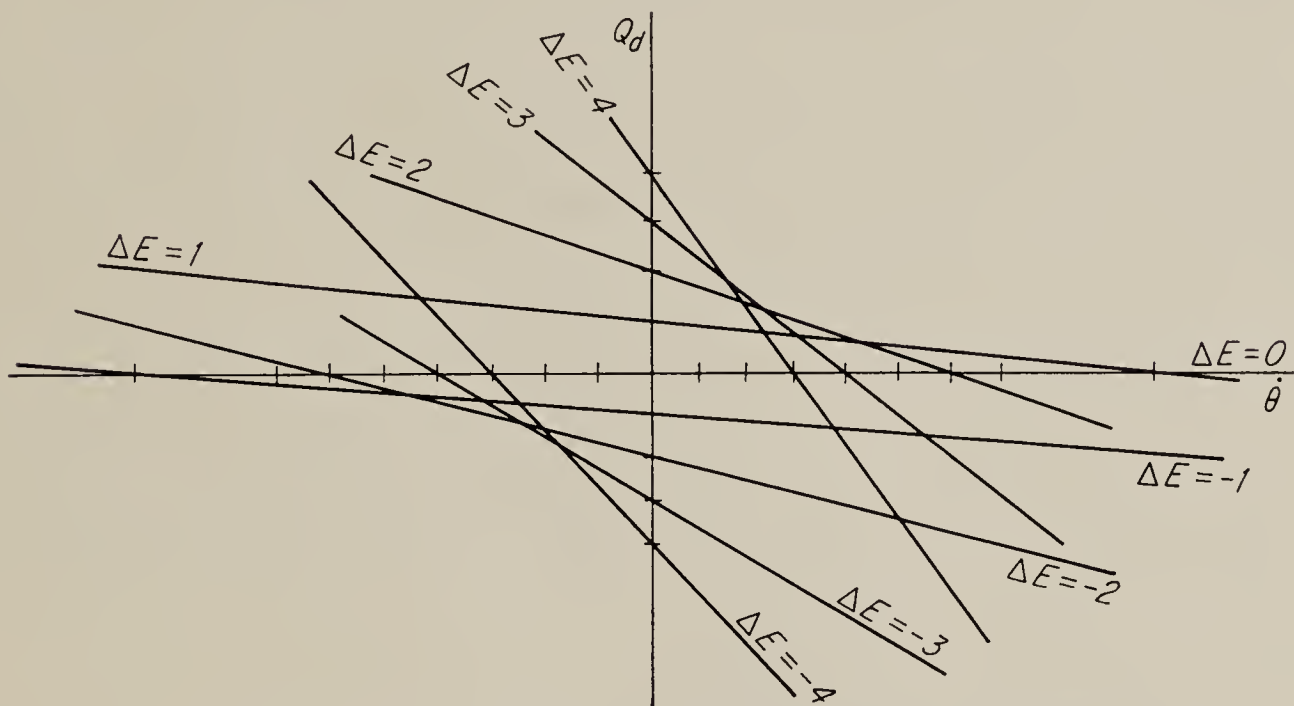


FIG. 4.27. Speed-torque curves for split-field series motor.

respectively. Inserting these approximations into Eq. (4.77) and equating to Eq. (4.78) yield the desired result:

$$\hat{\theta} \approx \frac{[4k_t E_0 \Delta \hat{E} / R_f (R_f + 2R_a)] - \hat{Q}_L}{s(Js + B)} \quad (4.79)$$

We see that, under the assumptions made, the motor does not supply electrical damping and that the only damping is due to mechanical friction. As a matter of fact, in the absence of friction the motor shaft position is the double integral of ΔE , the control voltage. This behavior differs quite markedly from that observed in the separately excited motor discussed in the previous sections.

It should be noted, however, that in general the damping supplied is proportional to the negative of the slope of the speed-torque curve (see Sec. 4.16). Hence, at large speeds and large control voltages, the split-field series motor is very heavily damped. The implications of this statement and the effect of this nonlinear damping on a servo loop in which a split-field series motor is used will not, however, be pursued here. As

has already been mentioned, the motor is often used with a polarized relay in a circuit such as is shown in Fig. 4.28, where either one field or the other is excited. The combined nonlinearities of the relay and the

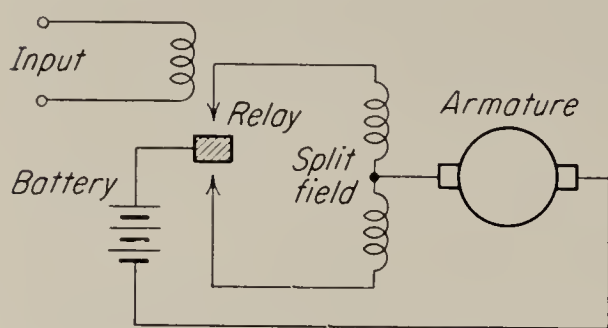


FIG. 4.28. Relay-operated split-field series motor.

motor are then usually such that only the most approximate analysis is possible,¹ and the servo must be designed more by experience and trial and error than by any sort of exact analysis.

4.20. The Field-controlled Motor. Still another motor-control method used in servo systems employs the field-controlled motor, in which the armature is supplied with a constant current. Since the control amplifier needs to supply only the relatively small field current, this drive may be used in moderately large power applications without including a generator in the loop to provide additional power amplification. In a sense the motor acts as a power amplifier itself. A generator is, however, needed to supply the required constant armature current; hence the amount of equipment needed is still about the same as for a Ward-Leonard system of similar power rating. The only advantage from the point of view of installed power capacity is that the constant current might possibly be furnished directly from the a-c line by a judicious arrangement of transformers and rectifiers. This equipment might be lighter, less expensive, and easier to maintain than a rotating generator.

The motor is commonly built with a center-tapped field coil to permit direct operation by a push-pull amplifier. Since it has already been shown in Sec. 4.2 that a push-pull field can always be replaced by an equivalent single-ended field, we consider this type here, as shown in Fig. 4.29. If linearity of the magnetic path is assumed, the developed torque is given by

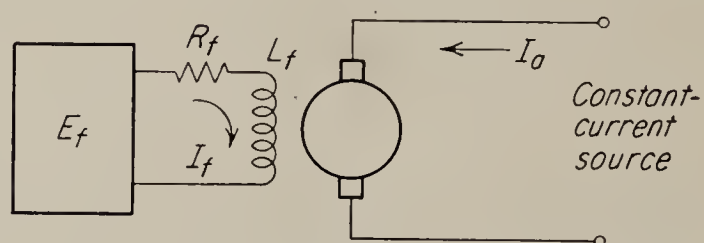


FIG. 4.29. A schematic diagram for the field-controlled d-c motor.

$$\hat{Q}_d = k_t I_a \hat{i}_f \quad (4.80)$$

The field current is given by

$$\hat{i}_f = \frac{\hat{E}_f}{R_f(1 + T_f s)} \quad (4.81)$$

where $T_f = L_f/R_f$.

If the armature has a moment of inertia J and a coefficient of viscous

¹ R. J. Kochenburger, A Frequency Response Method for Analyzing and Synthesizing Contactor Servomechanisms, *Trans. AIEE*, vol. 69, pp. 270-284, 1950.

friction B and is connected to a load requiring a torque \hat{Q}_L , the torque balance shown in Eq. (4.78) applies. Hence, the transfer function is

$$\hat{\theta} = \frac{(k_t I_a / R_f) \hat{E}_f}{s(T_f s + 1)(Js + B)} - \frac{\hat{Q}_L}{s(Js + B)} \quad (4.82)$$

Here again we note that the motor does not provide any electrical damping; the only damping is due to mechanical friction. As in the split-field series motor, if friction is negligible, the motor acts as a double integrator.

Very often the armature current for this type of motor is not obtained from a true constant-current source, but rather an approximately constant current is obtained by connecting the armature to a constant-voltage supply through a high resistance. If the motor is permitted to run only at speeds low enough that the counter emf remains negligibly small compared to the supply voltage, the armature current will remain almost constant. This method of supplying the armature has the advantage of doing away with an expensive and complicated constant-current supply. A large amount of power is, of course, wasted in the resistance, but since the drive can be used only for power outputs of about 500 watts or less, this is a relatively minor disadvantage.

If the voltage supplying the armature is V and if the total impedance in the armature circuit (including series resistance, armature resistance, and inductance) is $R + j\omega L$, the armature current is given by

$$i_a = \frac{V - E_c}{R + j\omega L} \quad (4.83)$$

where E_c , the counter emf, is given by

$$E_c = k_v i_f \dot{\theta} \quad (4.84)$$

The developed torque is still given by

$$Q_d = k_t i_a i_f \quad (4.85)$$

Note that Eqs. (4.83) through (4.85) are not written in Laplace transform notation, since they involve, on the right side, a product of two variables. In order to get an approximate result, we make a linear approximation of the two nonlinear equations as follows:

For small variations of i_f and $\dot{\theta}$ around some operating value, Eq. (4.84) may be approximated by the linear term of its Taylor expansion

$$\begin{aligned} \Delta E_c &= \left. \frac{\partial E_c}{\partial i_f} \right|_{E_c=E_{co}} \Delta i_f + \left. \frac{\partial E_c}{\partial \dot{\theta}} \right|_{E_c=E_{co}} \Delta \dot{\theta} \\ &= k_v \dot{\theta}_o \Delta i_f + k_v i_{fo} \Delta \dot{\theta} \end{aligned} \quad (4.86)$$

where E_{co} , $\dot{\theta}_o$, and i_{fo} are the operating-point values of the quantities in question. Similarly, Eq. (4.85) becomes

$$\Delta Q_d = k_t i_{ao} \Delta i_f + k_t i_{fo} \Delta i_a \quad (4.87)$$

If we consider only small variations of i_a and E_c , Eq. (4.83) becomes

$$\Delta \hat{i}_a = \frac{-\Delta \hat{E}_c}{R + Ls} \quad (4.88)$$

since V is a constant. If now Eqs. (4.86), (4.87), and (4.88) are combined with the torque balance [Eq. (4.78)] we obtain

$$\Delta \hat{\theta} = \frac{\Delta \hat{i}_f [i_{ao}(R + Ls)/k_v i_{fo}^2 - \dot{\theta}_o/i_{fo}] - \Delta \hat{Q}_L(R + Ls)/k_v k_t i_{fo}^2}{s[J s(R + Ls)/k_v k_t i_{fo}^2 + B(R + Ls)/k_v k_t i_{fo}^2 + 1]} \quad (4.89)$$

If we let $R + Ls = Z$ we note that this expression is not much different from the expression for $\hat{\theta}$ in the armature-controlled motor [Eq. (4.61)]. In fact, except for the “gain constant” term multiplying $\Delta \hat{i}_f$, the difference consists in the substitution of $k_v k_t i_{fo}^2$ for $k_v k_t$. Thus when the motor operates around an operating point other than the standstill point where i_{fo} and $\dot{\theta}_o$ are zero, its characteristics would be expected to be quite similar to those of the armature-controlled motor. Note, however, that if $i_{fo} = \dot{\theta}_o = 0$, the transfer function becomes

$$\Delta \hat{\theta} = \frac{k i_{ao} \Delta \hat{i}_f - \Delta \hat{Q}_L}{s(Js + B)} \quad (4.90)$$

Under this operating condition the motor loses all electrical damping and becomes a double integrator if B , the coefficient of friction, is negligibly small. Note that Eq. (4.90) is essentially the same as Eq. (4.82); therefore one may conclude that the motor characteristics for operation about zero field current are not affected by the method used for exciting the armature. It is also interesting to note that the speed-torque curves of this motor have exactly the same form as those plotted in Fig. 4.27 for the split-field series motor. Thus the characteristics of the two motor types should be identical, and conclusions reached by analyzing one should be applicable to the other.

PROBLEMS

4.1. Obtain the transfer function of the Regulex generator system (Fig. 4.15) and determine the value of R_1 required for critical tuning. The slope of the magnetization curve of the pilot generator is k_1 volts/amp of field current, while the slope of the magnetization curve of the main generator is k_2 volts/amp.

4.2. Show in what way the transfer functions of the Amplidyne are affected by the removal of the compensating winding in the direct-axis circuit. Indicate in what way the machine becomes a constant-current Rosenberg generator, and find the transfer function relating control-field current to load current.

4.3. The Amplidyne shown in the equivalent circuit of Fig. 4.20 is loaded by a pure resistance R_o . Assuming that the inductance L_d is negligible, find the transfer function \hat{V}_o/\hat{E}_f . Note that the time constant T_q does not appear in this transfer function.

4.4. The Amplidyne shown in the equivalent circuit of Fig. 4.20 drives a d-c motor having a moment of inertia J , armature resistance R_m , a counter-emf constant k_v , and a torque constant k_t . The inductance of the armature and the mechanical fric-

tion of the armature are negligible. Also neglect the inductance L_d of the Amplidyne. Find the transfer function $\hat{\theta}/\hat{E}_f$, where θ is the motor shaft position. Note that the time constant T_q does not appear in this transfer function.

4.5. The following two tests are performed on an Amplidyne to determine its characteristics:

a. Open-circuit step-function test: For a step function of field current of 1 ma the open-circuit output voltage is $E_o = 10(1 - e^{-50t})$ volts.

b. Short-circuit step-function test: For a step function of field current of 1 ma the short-circuit output current is $I_o = 1(1 - e^{-500t})$ amp.

This Amplidyne is used to drive a d-c shunt motor which has an armature resistance of 1 ohm, $k_v = 0.1$, $k_t = 0.1$, $J = 0.0002$ in a consistent set of units; friction, armature inductance, etc., are negligible.

Assume the simplest possible equivalent circuit for the Amplidyne; i.e., all mutual inductances are zero, and the inductance of the direct axis is negligible. Find a relation for the transfer function from Amplidyne field current to motor speed.

4.6. A d-c motor has the following ratings:

No-load speed	$= \omega_{NL}$	radians/sec
Full-load speed	$= \omega_{FL}$	radians/sec
No-load current	$= I_{NL}$	amp
Full-load current	$= I_{FL}$	amp

a. Find the ratio BR/k_vk_t in terms of these ratings, where B = coefficient of viscous friction, R = armature resistance, k_v = voltage constant, k_t = torque constant.

b. Determine this ratio for a motor for which

No-load speed	$= 1,020$ rpm
Full-load speed	$= 1,000$ rpm
No-load current	$= 0.2$ amp
Full-load current	$= 4.5$ amp

4.7. Show that the characteristic speed-torque curves of the field-controlled d-c motor shown in Fig. 4.30 are similar to the speed-torque curves of the split-field series motor (Fig. 4.27). Show that the quasi-linear transfer function of the motor is, therefore, similar to that of the split-field series motor.

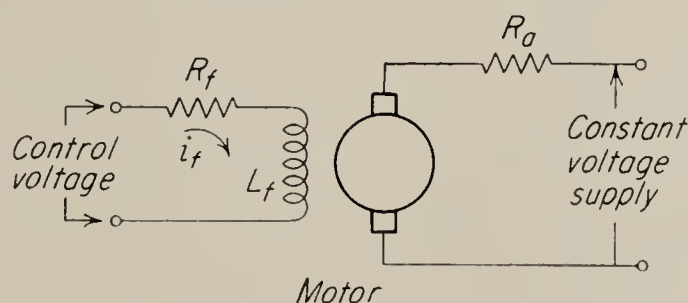


FIG. 4.30

CHAPTER 5

SYNCHROS AND RELATED DEVICES

5.1. Accuracy and Precision. An important component of almost all control systems is the device that converts the input and output variables from their original physical form to a form that is more convenient to use. Devices of this type are referred to as *pickup devices*, or *transducers*. In an electromechanical system such transducers convert the position or velocity of the output into electric signals, which are more conveniently handled in amplifiers and networks. Since servo systems are used to control all sorts of physical quantities, there is a large variety of transducers of different designs and for different applications, and it would be possible to write several volumes on the subject.¹ Space limitations here permit us, however, to discuss only a small segment of this large topic, and we shall confine our attention primarily to a small number of the more important electromechanical transducers, since these illustrate the basic

problems involved and the methods of attack.

The primary requirement to be fulfilled by a transducer is accuracy. Since the transducer always functions outside of the control loop, its inaccuracies cannot be compensated by means of feedback. A servo cannot be any more accurate than its transducer. The *error* in a transducer may be thought of as consisting of two components: a *repeatable* com-

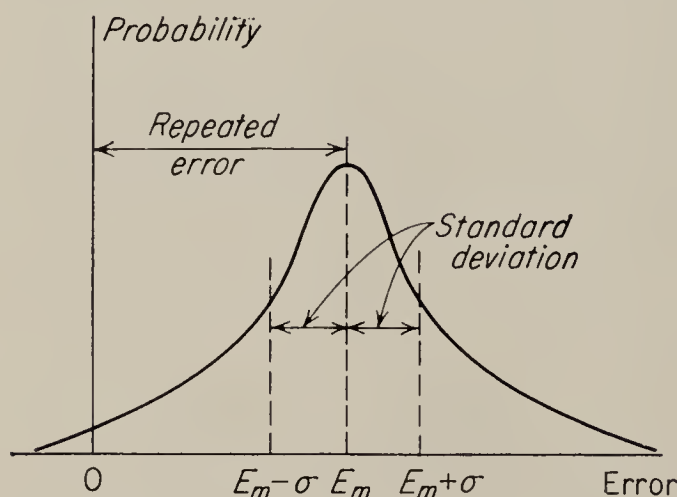


FIG. 5.1. Normal error curve.

ponent and a *random* component. In the absence of any direct knowledge concerning the randomness of the error, it is usual to assume that the probability of the error follows the so-called *normal law*:

$$p(E) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(E-E_m)^2/2\sigma^2} \quad (5.1)$$

This is a bell-shaped function, shown in Fig. 5.1. We define the repeat-

¹ See, for instance, Draper, McKay, and Lees, "Instrument Engineering," vols. 1-3, McGraw-Hill Book Company, Inc., New York, 1952-1955.

able component of the error as the average value E_m , and although the probability curve indicates that the random error extends over a very large range (theoretically infinite), the randomness is usually measured by σ , the standard deviation defined in Eq. (5.1). It is convenient to normalize these errors with respect to the range of the measured variable and to agree on the following definitions:

$$\text{Accuracy} \triangleq \frac{\text{range of variable}}{E_m} \quad (5.2)$$

$$\text{Precision} \triangleq \frac{\text{range of variable}}{\sigma} \quad (5.3)$$

The accuracy and precision as defined in these equations increase for the better instrument, which is in accordance with normal usage of these words. Note that the error contributing to poor accuracy is in the nature of an error in calibration. It is normally desirable that transducers be linear, i.e., that the output variable be directly proportional to the input variable. Any departure from linearity would therefore give rise to a repeatable error in a nominally linear transducer and thus reduce the accuracy as here defined. Also, in many servomechanism applications, two transducers with closely matched characteristics are used, one at the input and the other at the output. Any mismatch between these transducers would result in a repeatable error and adversely affect the accuracy of the system. In the following discussion of errors in transducers we shall be concerned primarily with repeatable errors, since the random errors contributing to lack of precision are primarily a function of the workmanship and care expended in the construction of the transducer.

In addition to accuracy and precision, another desirable characteristic of transducers is high input impedance; i.e., the transducer should require a minimum amount of power for its operation. This is particularly important in transducers operating at the input to a servomechanism, since one of the primary functions of an automatic control system is accurate power amplification between input and output. In some cases, it is required that only an infinitesimally small amount of power be extracted from the input; photoelectric devices are often used in such applications. Other desirable characteristics are rapidity of response and low noise output.

5.2. Elementary Operation of Synchros. The most common form of electromagnetic transducer used to convert an angular position of a rotating shaft into an electric signal is the *synchro* (also called *selsyn* or *autosyn*). It is typical of the class of induction transducers commonly employed in control and measurement applications, and the results of the discussion on synchros are applicable with some modifications to these other transducers also.

In servo applications, synchros are commonly employed in pairs in an arrangement similar to the one shown in Fig. 5.2. Two different types of synchros are used, the *generator* (also referred to as *transmitter*) and the *control transformer*. Other synchro devices include *motors* (*receivers* or *repeaters*) and *differential units*. These will be discussed later. In the system shown, the output voltage from the control transformer is used as an indication of the relative shaft positions θ and α of the two machines. Thus, the system acts as a comparator of the two shaft positions and as a transducer converting this difference of shaft positions into an electric signal.

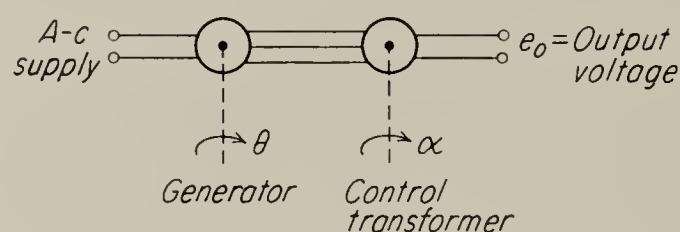


FIG. 5.2. Schematic of a synchro pair.

The construction of a synchro is very similar to that of a miniature, three-phase synchronous alternator. The stator is constructed with standard slotted steel punchings containing a three-phase winding that is

usually Y-connected. In order to minimize slot effects, the stator slots are often skewed. There are three different types of rotors in common use: the *salient-pole* type, the *umbrella* type, and the *cylindrical*, or *drum*, type (see Fig. 5.3). All of these rotors are wound for two electric poles, and the windings are brought out to slip rings so that continuous rotation is possible. Salient-pole rotors are used for synchro generators or motors. Both the other types of rotors are used in control transformers, where it is desirable to have uniform reluctance all the way around the air gap. Both the stator and rotor windings of control trans-

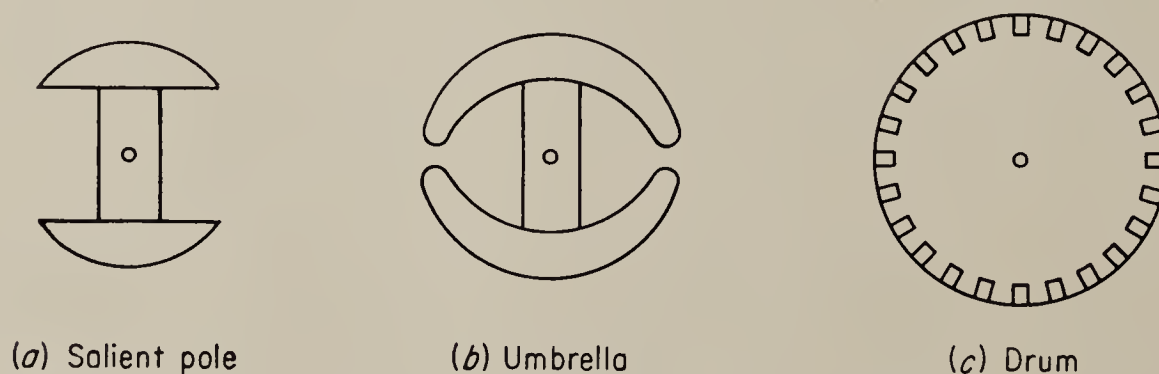


FIG. 5.3. Typical synchro-rotor cross sections.

formers usually have a higher impedance than those of the generator so that it is possible to excite several control transformers from one generator.

An elementary explanation of the operation of a synchro pair consisting of a generator and control transformer may be obtained by reference to Fig. 5.4. The alternating current flowing in the generator rotor sets up an alternating flux, as shown in the figure, and produces alternating voltages in the stator by ordinary transformer action. These voltages cause currents, all of the same phase but of different magnitudes, to circulate

in the stator windings. The result is that an alternating flux pattern is set up in the stator of the control transformer; this pattern has ideally the same space position as that in the generator. If the rotor of the control transformer is turned so that its pole axis is at right angles to this flux, the output voltage e_o is zero. However, in general, if the reluctance of the air gap of the control transformer is independent of rotor position, the output voltage is an a-c voltage proportional to the cosine of the angular difference between the pole axes of the rotors of the generator and control

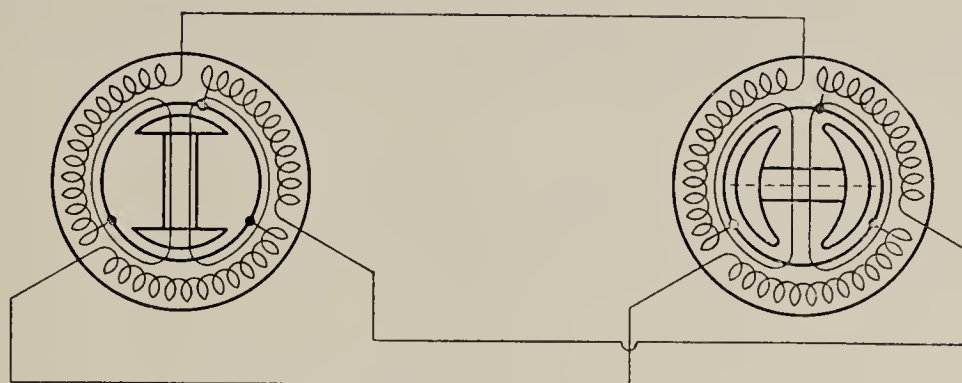


FIG. 5.4. Schematic of synchro pair showing stator connections.

transformer. Thus, if the voltage applied to the generator rotor is $E \sin \omega t$, then the control-transformer output voltage is (see Fig. 56)

$$e_o = E_{om} \cos (\theta - \alpha) \sin (\omega t - \beta) \quad (5.4)$$

where θ and α are the angles, measured from the same reference, of the pole axes of the generator and control-transformer rotors, respectively. E_{om} is the maximum value of the output voltage. It normally is equal to $\sqrt{2} \times 57$ volts in 115-volt synchros. The phase shift β between input and output voltages is the sort of phase shift found between primary and secondary voltages in a transformer and is due to the resistance of the windings. Since the ratio of resistance to reactance is usually quite small, β also is usually of the order of only a few degrees and is often neglected.

In operation in a servo system the rotor of one of the synchros, usually the control transformer, is connected to the output shaft, and the input is applied to the other rotor. The voltage e_o is applied to the amplifier, which feeds the motor and turns the output shaft. Hence the servo operates to make the output voltage from the control transformer zero, so that, under quiescent conditions, $\theta - \alpha = 90^\circ$. It is convenient, therefore, to define an angle $\delta = \alpha + 90^\circ$, in order that the condition for zero output voltage may be $\delta = \theta$. In this case

$$e_o = E_{om} \sin (\delta - \theta) \sin (\omega t - \beta) \approx E_{om}(\delta - \theta) \sin (\omega t - \beta) \quad (5.5)$$

if $\delta - \theta$ is a small angle. The output voltage is seen to be an alternating voltage whose amplitude is proportional to the angular difference between the rotor shafts of the two synchros, while the output phase angle depends

on the sign of this difference. Thus, in order to recover the complete information contained in the output signal, it must be passed through a phase-sensitive detector, or discriminator. Devices of this sort are discussed in the next chapter.

It should be noted that, despite its appearance, a synchro is a single-phase device, and all voltages and currents measured throughout a synchro system are essentially in phase except for small, incidental phase shifts such as the angle β mentioned above.

Thus far we have tacitly assumed δ and θ to be constant in time. Suppose now, however, that either δ or θ or both are oscillated sinusoidally with a relatively low frequency of ω_s radians/sec. Thus

$$\delta - \theta = \sigma_M \sin \omega_s t$$

If σ_M is small, the approximate form of (5.5) applies, and

$$e_o \approx E_{om} \sigma_M \sin \omega_s t \sin (\omega t - \beta) \quad (5.6)$$

This output voltage is shown in Fig. 5.5 together with the voltage applied to the generator, the so-called *reference* voltage. The phase angle β

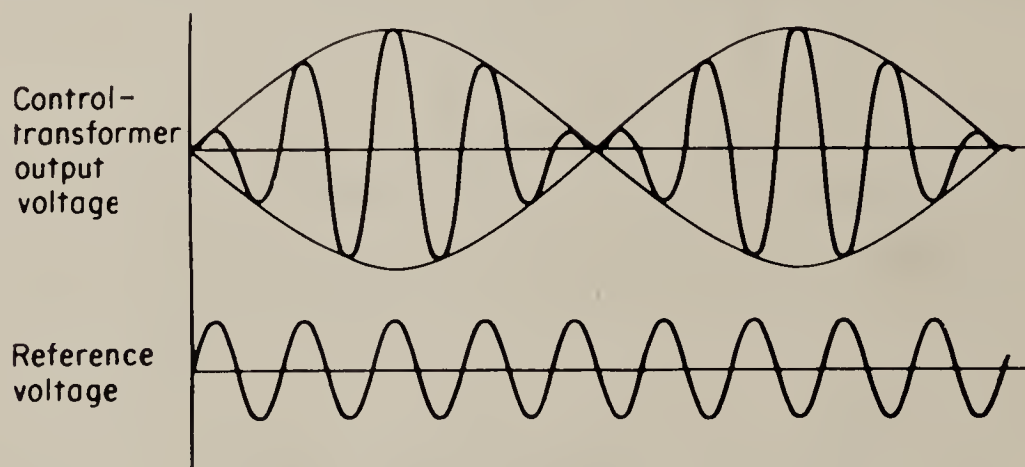


FIG. 5.5. Control-transformer output-voltage waveforms.

is taken to be zero. The figure shows the 180° phase reversal of the output voltage relative to the reference when the difference $\delta - \theta$ becomes negative. The output voltage is an example of a suppressed-carrier amplitude modulation. This may be shown by writing equation (5.6) in the equivalent form:

$$e_o = \frac{E_{om} \sigma_M}{2} \{ \cos [(\omega - \omega_s)t - \beta] - \cos [(\omega + \omega_s)t - \beta] \} \quad (5.7)$$

The voltage is seen to be composed of two sidebands having the frequencies $\omega - \omega_s$ and $\omega + \omega_s$, but there is no carrier term (ω) present. The form of this voltage is typical of the outputs of a large number of transducers used in servomechanisms in which the input variable is used to modulate the amplitude of an a-c carrier.

5.3. Static Errors and Residual Voltages. The relationship between the rotor positions and control-transformer output voltage as given in the previous section describes an ideal situation, from which all synchros depart to a certain, even if usually very small, extent. It is found, for instance, that, as the rotors of the two synchros approach correspondence, the output voltage does not go to zero, as Eq. (5.5) indicates, but instead only goes to a minimum. The residual voltage remaining at the minimum consists primarily of a component at carrier frequency ω but 90° out of phase with the predominant phase of the output signal for large differences in rotor displacement. It is therefore referred to as the *quadrature voltage*, or *quadrature error*.

It can be shown that one reason for the existence of the quadrature error is that the impedances making up the stator circuit do not all have the same phase angle. Thus consider the system shown in Fig. 5.6. The

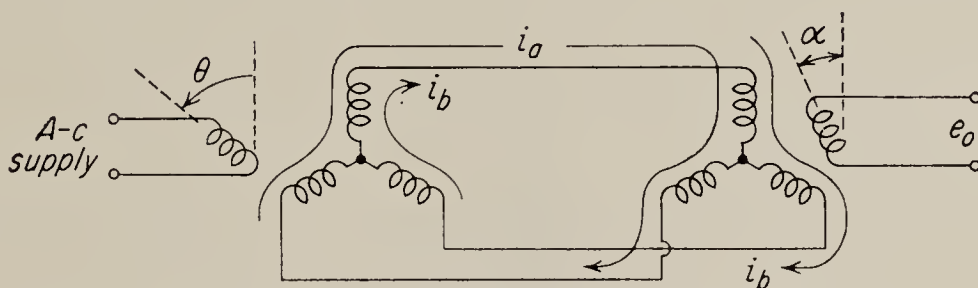


FIG. 5.6. Currents and voltages in a synchro system.

two currents i_a and i_b are alternating at the carrier frequency ω . Their magnitude depends on θ , the rotor angle of the generator, and although ideally they are in time phase, in practice a small difference in the phase angles of the impedances of the stator results in a small phase angle γ between the currents. Thus, let

$$\begin{aligned} i_a &= I_a \sin \omega t \\ i_b &= I_b \sin (\omega t - \gamma) \end{aligned} \quad (5.8)$$

The control-transformer output voltage is produced by the mutual inductance between the control-transformer stator and rotor windings. This mutual inductance is a periodic function of α , the rotor angle. For the usual cylindrical rotor construction this inductance may be assumed to be a sinusoidal function of α . In order to determine the exact form of this inductance, we note from Fig. 5.6 that, when $\alpha = 30^\circ$, the coupling between the stator circuit traversed by current i_b and the rotor is a maximum. Similarly, when $\alpha = -30^\circ$, the coupling between i_a and the rotor is maximum. Hence, if we let M_c be the maximum value of the mutual inductance between either i_a or i_b and the rotor, then the output voltage may be written

$$e_o = M_c \cos (\alpha + 30^\circ) \frac{di_a}{dt} + M_c \cos (\alpha - 30^\circ) \frac{di_b}{dt} \quad (5.9)$$

By the use of Eq. (5.8), Eq. (5.9) becomes

$$e_o = \omega M_c [I_a \cos (\alpha + 30^\circ) \cos \omega t + I_b \cos (\alpha - 30^\circ) \cos (\omega t - \gamma)] \quad (5.10)$$

By expanding the $\cos (\omega t - \gamma)$ term, this becomes

$$e_o = \omega M_c \{ [I_a \cos (\alpha + 30^\circ) + I_b \cos \gamma \cos (\alpha - 30^\circ)] \cos \omega t + [I_b \sin \gamma \cos (\alpha - 30^\circ)] \sin \omega t \} \quad (5.11)$$

This equation indicates that the rms value of the control-transformer output cannot go to zero for any value of α , unless I_b is zero, a condition occurring for only two positions of the generator rotor. The second term in the equation is the quadrature error referred to above, and for small γ it is approximately proportional to γ , the phase difference between the stator currents. It is clear that this voltage can be kept to a minimum by making certain that the impedances of the stator circuit all have the same phase angle. Usually synchros are wound quite carefully, and the windings are almost identical, but unless the connecting cable is perfectly constructed, it may introduce unbalancing impedance that will give rise to considerable error.

In addition to the effect discussed above, one usually finds harmonic voltage components near the synchro nulls. Some of these are due to harmonics present in the supply voltage, but most of them are due to nonlinearities in the iron, etc. Quadrature voltages of fundamental frequency are also sometimes caused by unbalanced capacitive coupling between stator and rotor. This effect is usually noticeable only at higher carrier frequencies, and although not important in synchros, it is an important cause of quadrature error in many other suppressed-carrier types of transducers.

The phase discriminator, which in one form or another is always used to recover the actual error information from the synchro signal, can, by proper adjustment, be made insensitive to all quadrature and harmonic voltages. Hence these residual effects do not actually result in any appreciable servo error. However, if it is desired to amplify the synchro error signal in an a-c amplifier, this amplifier must be designed to handle input amplitudes that are at least as large as the maximum quadrature signal, and preferably somewhat larger. Hence a large amount of quadrature voltage limits the amount of a-c gain that can be used ahead of the discriminator. If more gain is needed, a d-c amplifier can be used after the discriminator. However, since d-c amplifiers tend to drift, it is usually desirable to have as much of the required gain as possible in the a-c amplifier. For this reason it is important to reduce the quadrature error to a minimum. Another disadvantage of a large quadrature signal is that it results in a considerably larger ripple output from the discriminator (see Chap: 6) and therefore requires more thorough filtering.

The component of primary importance in the control-transformer output voltage is the inphase component given by the coefficient of $\cos \omega t$ in Eq. (5.11). When a synchro pair is used in a servo, the servo system acts to make this term go to zero. Thus, if zero voltage does not correspond to exactly the same angular difference between the rotors under all operating conditions, an error is introduced into the system. In practice, two types of errors are found: (1) *static errors* depending on the position of the input and output shafts and (2) *velocity errors*, which are a function of the speed of the rotors.

A typical test record of the static error of a generator-control-transformer system as a function of generator rotor position is shown in Fig. 5.7. It shows that the error consists predominantly of a second- and

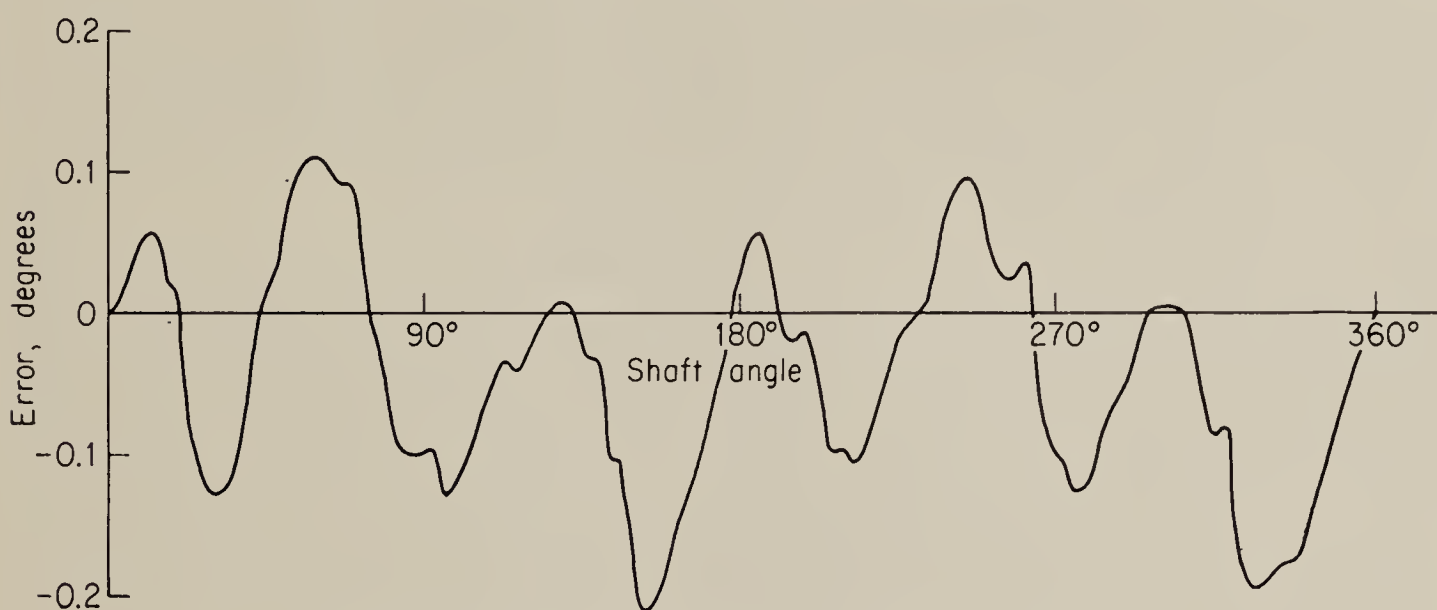


FIG. 5.7. Synchro error for zero output voltage.

sixth-harmonic function of rotor position. It can be shown that the second harmonic is caused by unbalance of the stator impedances, eccentricities of the generator rotor relative to the stator bore, or ellipticity of either the stator bore or the rotor surface. In fact, any sort of stator-circuit unbalance that might cause the maximum values or the phases of the stator currents I_a and I_b to be different can be the cause. For zero inphase output voltage we have from Eq. (5.11)

$$I_a \cos (\alpha + 30^\circ) + I_b \cos \gamma \cos (\alpha - 30^\circ) = 0 \quad (5.12)$$

I_a and I_b , the amplitudes of the two stator currents i_a and i_b shown in Fig. 5.6, are periodic functions of θ (a "negative amplitude" refers to a voltage 180° out of phase with one having a positive amplitude). The salient rotor construction commonly used in synchro generators may be expected to make I_a and I_b nonsinusoidal functions of θ , but for the present it is convenient to ignore this fact and to assume a sinusoidal variation. Inspection of Fig. 5.6 indicates that, with $\theta = 30^\circ$, $I_a = 0$ and, with $\theta = -30^\circ$, $I_b = 0$. Furthermore both I_a and I_b are positive as shown, when $\theta = 0^\circ$. Hence, if we let I_{am} and I_{bm} be the maximum values of I_a

and I_b , respectively, then on the assumption of a sinusoidal variation we have

$$\begin{aligned} I_a &= I_{am} \cos (\theta + 60^\circ) \\ I_b &= I_{bm} \cos (\theta - 60^\circ) \end{aligned} \quad (5.13)$$

If these values of I_a and I_b are inserted into Eq. (5.12), the condition for zero output voltage becomes, after some trigonometric and algebraic reduction,

$$\cos \left\{ \theta - \alpha - \tan^{-1} \frac{\left(\frac{I_{bm}}{I_{am}} \cos \gamma - 1 \right) \left(\frac{1}{2\sqrt{3}} - \frac{1}{\sqrt{3}} \cos 2\theta \right)}{1 + \left(\frac{I_{bm}}{I_{am}} \cos \gamma - 1 \right) \left(\frac{1}{2} + \frac{1}{\sqrt{3}} \sin 2\theta \right)} \right\} \quad (5.14)$$

In arriving at this result, we have used the fact that $\sin (\theta + \alpha)$ may be written as $\sin [2\theta - (\theta - \alpha)]$. Note that, if the stator circuits are balanced, $I_{am} = I_{bm}$ and $\gamma = 0$; hence the relation reduces to the ideal

$$\begin{aligned} \cos (\theta - \alpha) &= 0 \\ \theta - \alpha &= 90^\circ \end{aligned} \quad (5.15)$$

or

However, if for any reason the stator circuits are not exactly balanced, either I_{am} differs from I_{bm} or γ has a nonzero value, or both, so that an additional angle, which by Eq. (5.14) is a function of 2θ , must be added in Eq. (5.15) to reduce the inphase voltage output of the control transformer to zero. This additional angle represents an error in the synchro system.

The sixth-harmonic error component indicated in Fig. 5.7 can be shown by a similar argument to be caused by the salient-pole construction of the generator rotor. This results in a nonsinusoidal variation of the currents I_a and I_b with rotor angle that may be expressed in a Fourier series. Owing to symmetry, only the odd harmonics will be present, and it can be shown¹ that the third, ninth, etc., harmonics cannot exist because of the balanced three-phase arrangement of the stator. Hence, if we consider harmonics up to and including the ninth, there will be present, in addition to the fundamental, only the fifth and seventh harmonics of θ . Thus, if it is assumed that the stator impedances are all balanced so that $I_{am} = I_{bm}$, we have

$$\begin{aligned} I_a &= I_1 \cos (\theta + 60^\circ) + I_5 \cos (5\theta - 60^\circ) + I_7 \cos (7\theta + 60^\circ) \\ I_b &= I_1 \cos (\theta - 60^\circ) + I_5 \cos (5\theta + 60^\circ) + I_7 \cos (7\theta - 60^\circ) \end{aligned} \quad (5.16)$$

Here I_1 , I_5 , and I_7 are the maximum amplitudes of the fundamental, fifth, and seventh harmonics, respectively, and are assumed to be the same for

¹ See, for instance, Kerchner and Corcoran, "Alternating Current Circuits," John Wiley & Sons, Inc., New York, 1938, pp. 268ff. The discussion given there is for a polyphase circuit, but a similar argument can be used in the single-phase synchro system to show that the third space harmonic of stator current cannot exist in the lines.

I_a and I_b . The phase angles for the fifth- and seventh-harmonic terms can be deduced from inspection of Fig. 5.6 and the facts that, for $\theta = 30^\circ$, $I_a = 0$ and, for $\theta = -30^\circ$, $I_b = 0$. Both I_5 and I_7 may be negative. If it is assumed that γ in Eq. (5.12) is zero and if Eq. (5.16) is then substituted into (5.12), there results, after some reduction, the new relation for zero inphase output:

$$I_1 \cos (\theta - \alpha) + I_5 \cos (5\theta + \alpha) + I_7 \cos (7\theta - \alpha) = 0 \quad (5.17)$$

But $\cos (5\theta + \alpha) = \cos [6\theta - (\theta - \alpha)]$, and

$$\cos (7\theta - \alpha) = \cos [6\theta + (\theta - \alpha)]$$

Using these results, we obtain after further manipulation

$$\cos \left\{ \theta - \alpha - \tan^{-1} \frac{[(I_5/I_1) - (I_7/I_1)] \sin 6\theta}{1 + [(I_5/I_1) + (I_7/I_1)] \cos 6\theta} \right\} = 0$$

or

$$\theta - \alpha = 90^\circ + \tan^{-1} \frac{[(I_5/I_1) - (I_7/I_1)] \sin 6\theta}{1 + [(I_5/I_1) + (I_7/I_1)] \cos 6\theta} \quad (5.18)$$

Thus the presence of a sixth-harmonic error is shown. In modern high-quality synchros this error may be made very small by so arranging the pitch and distribution factors of the stator winding that the fifth and seventh harmonics tend to be canceled. For details of this procedure the reader is referred to standard texts on a-c machinery.¹

It should be noted that the developments just carried out are intended primarily to furnish qualitative insight into the possible causes of synchro error, and the results obtained should be regarded as approximate. For more accurate consideration of these errors and for analyses leading to improved synchro design methods, the interested reader is referred to papers by Chestnut,² Kronacher,³ and Rosenbloom, Weiss, and Fried⁴ in which the causes of synchro errors are treated in considerable detail.

The additional space harmonics apparently present in the error curve shown in Fig. 5.7 are due primarily to winding irregularities and slot effects. The maximum error is seen to be less than 0.2° , and in some synchros designed for particularly high accuracy, errors have been reduced to less than 0.1° .

¹ Puchstein, Lloyd, and Conrad, "Alternating Current Machines," John Wiley & Sons, Inc., New York, 1954, pp. 133-141.

² H. Chestnut, Electrical Accuracy of Selsyn Generator-Control-transformer System, *Trans. AIEE*, vol. 65, pp. 570-576, 1946.

³ G. Kronacher, Static Accuracy Performance of the Selsyn Generator-Control-transformer System, *Trans. AIEE*, vol. 69, pp. 645-653, 1950.

⁴ J. H. Rosenbloom, G. H. Weiss, and B. D. Fried, "Linear Lumped Parameter Analysis of Synchros," U.S. Naval Ordnance Laboratory, White Oak, Md. This study is in nine parts, NAVORD Reports 1710, 2260, 2172, 2173, 2570, 2346, 2829, 2569, 3633, and is summarized in a final volume, "A Handbook for Synchro Systems," by G. H. Weiss and G. L. Beyer, Jr., NAVORD Rept. 3600, December, 1953.

5.4. Velocity Errors. In addition to the static errors discussed in the previous section, synchros are subject to an error that is due to rotational velocity. Thus, if the rotors of a generator and control transformer are aligned to produce essentially zero output and if the rotors are then rigidly coupled together, so that there can be no relative motion between them, and if then the two synchros are rotated together, a voltage will appear at the control-transformer terminals. This voltage will consist in part of components due to the static errors discussed above, but there is also a steady component that increases with speed. Hence if the two synchros are used in a servo system that tends to null the inphase component of output voltage, the output tends to run slightly behind the input when a constant-velocity input signal is applied. This velocity error is in addition to the error normally found in servo systems having only one integration.

Qualitatively speaking, it is not unreasonable that such an error should exist, for whenever one electric circuit moves with respect to another, speed voltages are induced. This effect is discussed in some detail in Sec. 4.4. A quantitative analysis of the phenomenon in synchros requires, however, a considerable background of advanced synchronous-motor theory, and even an approximate treatment of it is beyond the scope of this text. We present here, therefore, only the results of some of the quantitative studies that have been made.^{1,2} The theoretical results given here are derived on the basis that the static errors and null voltages considered in the previous sections are zero. The further important assumption is made that the impedance of the control transformer is very much higher than that of the generator. Under these conditions the magnitude of the control-transformer output voltage is given by

$$|E_o| = k_e \sqrt{\frac{1 + Q^2}{[1 - (1 - v^2)Q^2]^2 + 4Q^2}} \{[v \cos \sigma - (1 - v^2)Q \sin \sigma]^2 + \sin^2 \sigma\} \quad (5.19)$$

and the phase angle of the output voltage relative to β_0 , the phase angle at standstill, is given by

$$\beta - \beta_0 = \tan^{-1} \left(\frac{1}{Q} \right) \left(\frac{1 + (1 + v^2)Q^2}{1 + (1 - v^2)Q^2} \right) + \tan^{-1} \left(\frac{1}{v \cot \sigma - (1 - v^2)Q} \right) \quad (5.20)$$

where $k_e \triangleq$ synchro gain constant, equal to $0.707E_{om}$ if E_o represents rms value of output voltage

¹ G. H. Weiss, "Linear Lumped Parameter Analysis of Synchros, IX, Effects of Angular Velocity on a Simple Control System," NAVORD Rept. 3633, U.S. Naval Ordnance Laboratory, White Oak, Md., Feb. 8, 1954.

² Chestnut, *op. cit.*

$v \triangleq$ ratio of rotational velocity, radians/sec, to excitation frequency, radians/sec

$Q \triangleq$ ratio of reactance to resistance in stator of control transformer

$\sigma \triangleq$ angular displacement between the two rotor positions, i.e.,
 $\sigma = \theta - \delta = \theta - \alpha - 90^\circ$

We note that the rotational velocity of the two synchros enters the two expressions only in the form of v , the ratio between the velocity and the frequency. Hence it appears that the velocity error for a given speed may be reduced by operating the synchros at a higher frequency. The fact that the phase angle of the output voltage changes with velocity shows that both the inphase and quadrature components of the output are functions of velocity. However, for low velocities such that $1 \pm v^2 \approx 1$,

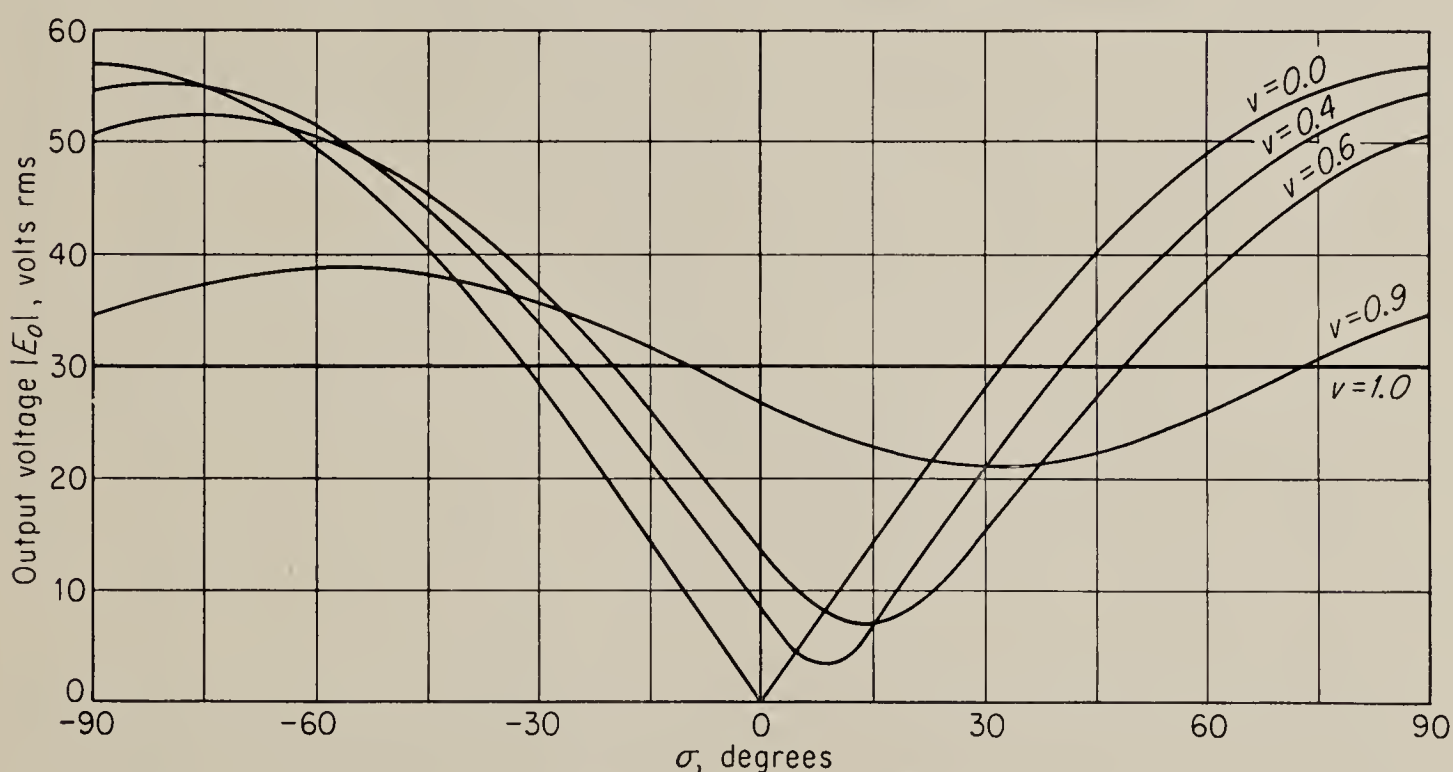


FIG. 5.8. Output voltage as a function of rotor displacement and velocity.

it may be seen that the phase shift is practically independent of velocity, and hence most of the output voltage consists of the inphase component.

The characteristics of Eqs. (5.19) and (5.20) are best displayed graphically,¹ as in Figs. 5.8 and 5.9. Q is taken equal to 3 and E_{om} to 57 volts. It is noted that for zero velocity the output voltage comes to a sharp null, and the phase angle changes abruptly from 0 to 180° at the point where the rotors are 90° apart, i.e., where $\sigma = 0$. As the velocity increases, the output voltage no longer goes to zero for any value of σ but only reaches a minimum, and this minimum shifts to positive values of σ . Of greater significance in a servo system using a discriminator that rejects the out-of-phase component of the output signal is the fact that the point at which the phase shift is 90° also shifts to positive values of σ with increased velocity. At this point the inphase component is zero, and hence the servo will be in error by at least the amount of this shift.

¹ Weiss, *op. cit.*

Comparison of Figs. 5.8 and 5.9 indicates that the point of minimum output voltage coincides quite closely with the point at which the phase shift is 90° , at least for relatively low velocity. Hence the error introduced into the servo system by the velocity effect is approximately equal

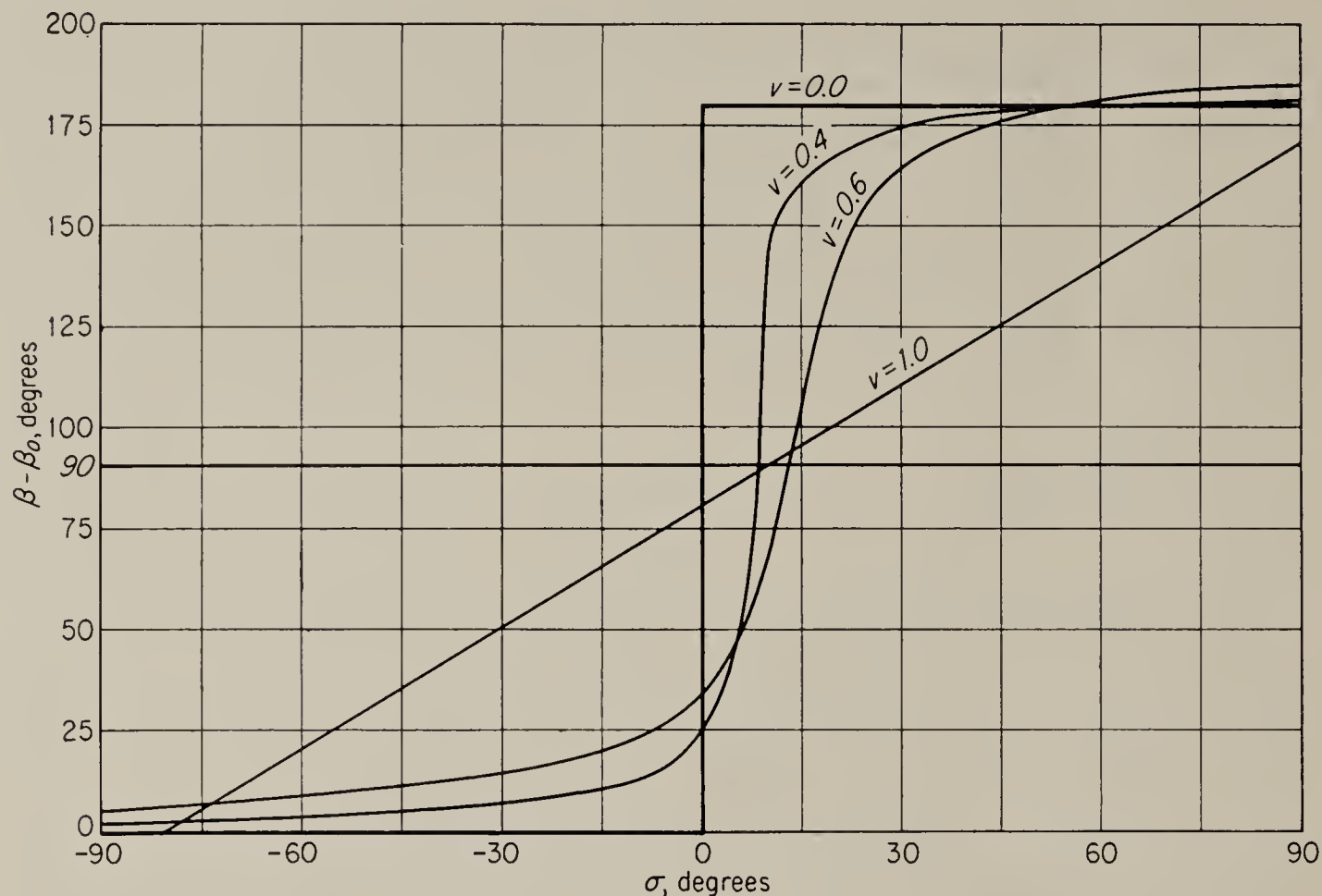


FIG. 5.9. Phase of output as a function of rotor displacement and velocity.

to the value of σ for which the output is minimum. This value, designated as σ_m , may be obtained by differentiating Eq. (5.19) with respect to σ and setting the derivative equal to zero. The result is given by

$$\tan 2\sigma_m = \frac{2vQ}{1 + (1 - v^2)Q^2} \quad (5.21)$$

and is plotted in Fig. 5.10 for various values of Q . For the relatively low velocities usually encountered, $1 - v^2 \approx 1$, and the error is small enough so that $\tan 2\sigma_m \approx 2\sigma_m$. Hence for low velocities Eq. (5.21) may be reduced to

$$\frac{\sigma_m}{v} \approx \frac{57Q}{1 + Q^2} \quad \text{degrees} \quad (5.22)$$

Thus in a 60-cps synchro system having $Q = 4$, the error is about $1^\circ/300$ rpm. Note that, by the definition of σ , $\sigma = \theta - \delta$. A positive σ means that θ is greater than δ ; hence for positive velocity the output shaft lags behind the input shaft.

The velocity error is, of course, present also whenever the synchro system is subjected to dynamic inputs of any sort and might be expected to

affect the synchro transfer function at high input frequencies. This transfer function cannot be obtained directly from Eqs. (5.19) and (5.20), since in deriving these equations steady-state conditions were assumed. In fact, a mathematical description of the complete dynamic situation, where both θ and α are allowed to vary in an arbitrary fashion, results in a set of differential equations with variable coefficients, and a simple transfer function in the usual sense does not, therefore, exist. A number of relatively simple results may, however, be obtained if the velocity of the generator rotor is assumed to be constant. Although no derivations

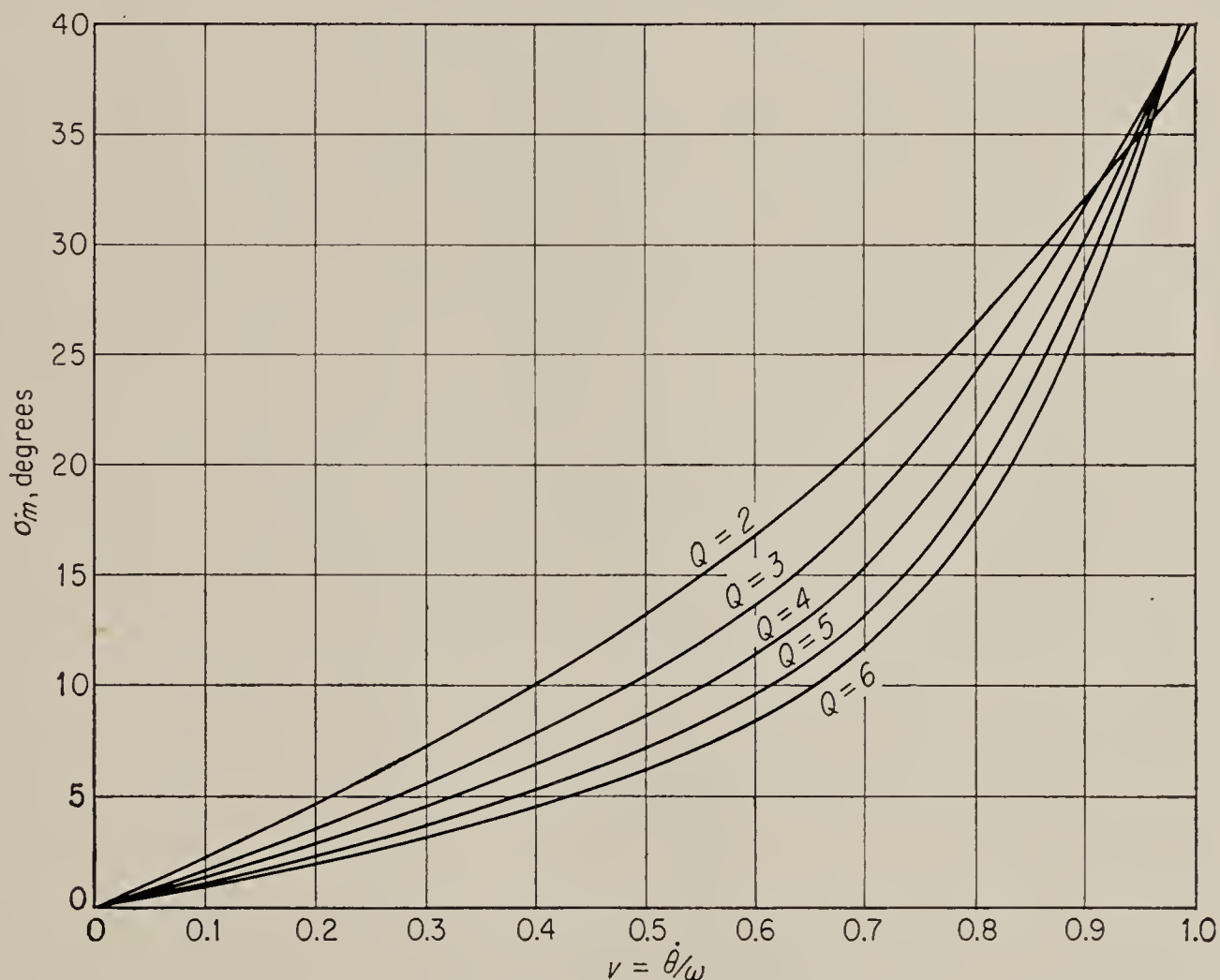


FIG. 5.10. Variation of σ_m as a function of velocity and Q [see Eq. (5.21)].

will be given here, it is possible to show¹ that under these conditions the output voltage is given by

$$E_o = \frac{k_e(Q - j)}{1 - (1 - v^2)Q^2 + 2jQ} \left\{ \frac{\dot{\sigma}}{\omega} [\cos \sigma + jQ(1 - v^2) \cos \sigma + jv \sin \sigma] + v \cos \sigma - Q(1 - v^2) \sin \sigma + j \sin \sigma \right\} \quad (5.23)$$

where ω is the line frequency in radians per second, the other symbols are as defined in connection with Eqs. (5.19) and (5.20), and $\dot{\sigma}$ is the difference between the velocities of the generator and control transformer. If $\dot{\sigma}$ is

¹ *Ibid.* Equation (5.23) is not given in Weiss's work but may be derived from the results given there.

zero, Eq. (5.23) can be reduced to a form equivalent to Eqs. (5.19) and (5.20). If $v = 0$, i.e., if the generator rotor does not rotate, Eq. (5.23) reduces to the simple expression

$$E_o = k_e \left(\sin \sigma - j \frac{\dot{\sigma}}{\omega} \cos \sigma \right) \quad (5.24)$$

For small values of σ this may be written approximately as

$$E_o \approx k_e \left(\sigma - j \frac{\dot{\sigma}}{\omega} \right) \quad (5.25)$$

It is clear that, although the velocity of the control-transformer rotor does contribute a component to the output, this component is 90° out of phase with the dominant term and would, therefore, be rejected by a properly adjusted phase discriminator following the control transformer. Hence, when the generator is not rotating, the transfer function relating discriminator output voltage to control-transformer rotor position is independent of frequency.

In general, the inphase component of the control-transformer output voltage does increase somewhat as the frequency of oscillation of either of the two rotors is increased. Equation (5.23) indicates, however, that the magnitude of the term contributed by the velocity becomes comparable to the magnitude of the term due to position only for velocities approaching synchronous velocity. Hence we infer that any time constants contributed by the velocity error to the synchro transfer function will be of the order of one period of the line frequency and will therefore be negligible.

An additional result deducible from Eqs. (5.23) and (5.24) is that, when σ is varied sinusoidally at relatively high frequencies, the output voltage

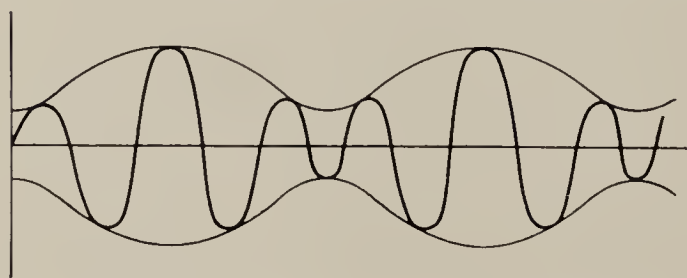


FIG. 5.11. Control-transformer output voltage for high-frequency signals.

from the control transformer will not have the idealized form shown in Fig. 5.5. The sharp null shown in that figure at points where $\sigma = 0$ will disappear, since $\dot{\sigma}$ is a maximum at these points. Hence the output voltage tends to be as shown in Fig. 5.11.

It is evident from Fig. 5.8 that the separation between maximum and minimum values of output voltage decreases with velocity. This may be interpreted as a reduction in gain $\partial E_o / \partial \sigma$ of the synchro system with velocity. Since we are again primarily concerned with the inphase component of the output, it is instructive to compute the ratio of a small change of inphase output voltage as a function of a small change in σ .

This result can be derived by use of Eqs. (5.19) and (5.20), and a simple expression results if this is evaluated for $\sigma = 0$:

$$\left. \frac{\partial[R_e(E_o)]}{\partial\sigma} \right|_{\sigma=0} = k_e \frac{Q^4(1 - v^2)^2 + 2Q^2 + 1}{Q^4(1 - v^2)^2 + 2Q^2(1 + v^2) + 1} \quad (5.26)$$

This expression is plotted in Fig. 5.12 for $Q = 3$; note that the change in gain is not very great until the velocity becomes greater than one-third of synchronous speed.

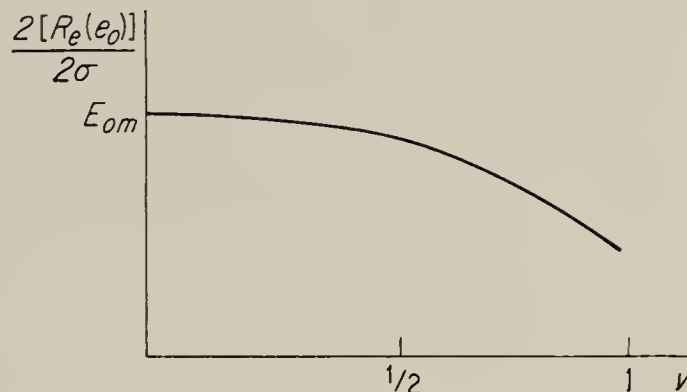


FIG. 5.12. Variation of gain as a function of speed.

5.5. Two-speed Systems. The static accuracy of servomechanisms using synchros can be made much higher than the static accuracy of the synchros themselves by the simple expedient of gearing up the rotor of the control transformer relative to the actual output. The

generator must, of course, be geared up in the same amount relative to the true input. Then if the static synchro error is ϵ , and if both synchros make N revolutions for each revolution of the output or input, the error at the output will be reduced to ϵ/N . This assumes, of course, that the gears do not introduce an additional error; but if precision gears are used, a considerable improvement in accuracy is possible. Commonly used values of N are 36, 31, and 25.

Unfortunately, a system in which the synchros are geared up by a factor of N will have $N - 1$ false points of synchronization. Thus, for instance, assume $N = 36$. Then for every 10° rotation of the output the control transformer rotates through 360° . Hence, if the system is in equilibrium for a given value of the input, it will also be in equilibrium when the output is displaced from its correct value by 10° , 20° , etc. This difficulty may be overcome by using a second set of synchros arranged to make one revolution for each revolution of the output or input. These synchros are referred to as the *one-speed synchros* and serve to bring the servo into approximate alignment. When the error in the one-speed system is sufficiently small, the one-speed system may be switched off and the *N-speed system* allowed to take over to complete the final accurate alignment of the over-all system. A system of this sort is shown schematically in Fig. 5.13. For the system to work properly it is, of course, necessary for the stators of the two pairs of synchros to be clamped into their supporting frames in such a way that the two synchros tend to line up the output to exactly the same point.

It should be noted that, in all cases where a two-speed synchro transmission is employed, the one-speed transmission is used only for approxi-

mate synchronization and is not operative during normal operation of the servo. Hence the performance characteristics of a servo considered in the design are always those observed with the high-speed system in control.

A number of circuits are available to perform the required switching between the high-speed and one-speed systems. One simple circuit utilizing neon tubes as switches is shown in Fig. 5.14. Its operation

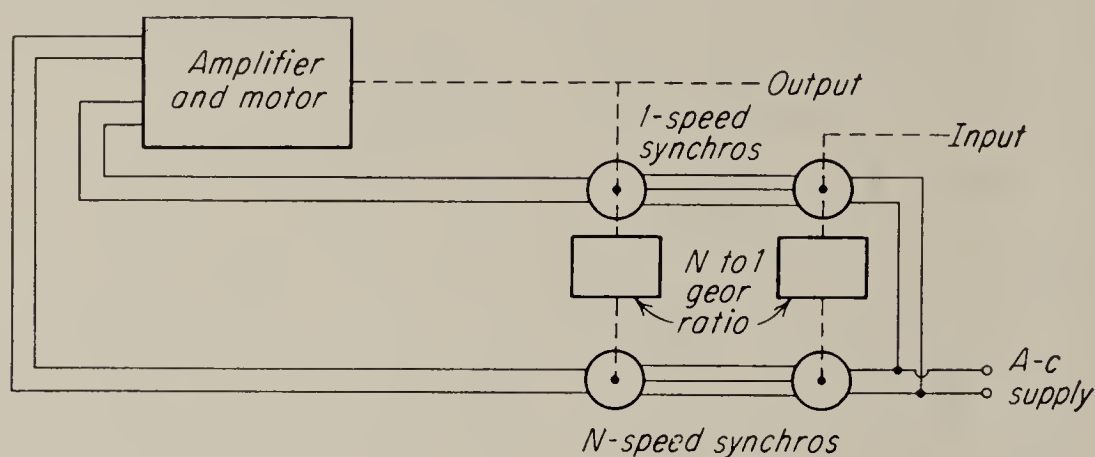


FIG. 5.13. Servo using two-speed synchro transmission.

depends on the fact that the amplifier has an infinite input impedance and may be explained as follows: The signal coming from the one-speed control transformer is amplified in V_1 and V_2 and is applied through the transformer T_2 to the neon tubes. If the signal is large enough, the neon tubes fire and provide a low-impedance path for the one-speed signal to the amplifier input. At the same time the signal from the high-speed control

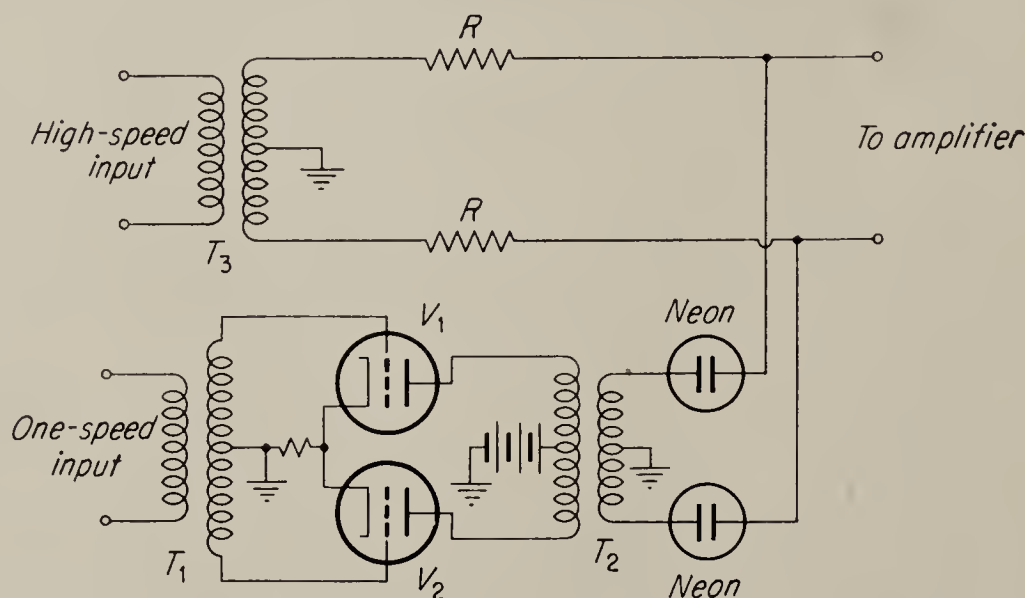


FIG. 5.14. Neon-tube switching circuit for two-speed synchro.

transformer is short-circuited by the neon tubes, since the resistors R have a very much higher resistance than the circuit containing the neon tubes. As the servo approaches synchronism, the signal from the one-speed control transformer decreases to such a value that the neon tubes go out. They now offer essentially infinite resistance and therefore disconnect the one-speed signal from the amplifier. The high-speed signal, however, now passes through the amplifier, since the resistors R are very much

smaller than the input impedance of the amplifier and the resistance of the neon-tube circuit. The turns ratio of the high-speed input transformer must be so chosen that the high-speed signal is never sufficiently large to fire the neon tubes.

For a given small rotation of the output away from synchronism, the output voltage of the N -speed synchro is N times as high as that of the one-speed synchro. Therefore, the loop gain with the high-speed channel operating is inherently N times as large as with the one-speed system switched on. An advantage of the circuit described here is that the extra amplification included in the one-speed channel permits the loop gain to remain approximately constant as the system is switched from one channel to the other. This contributes to smooth operation.

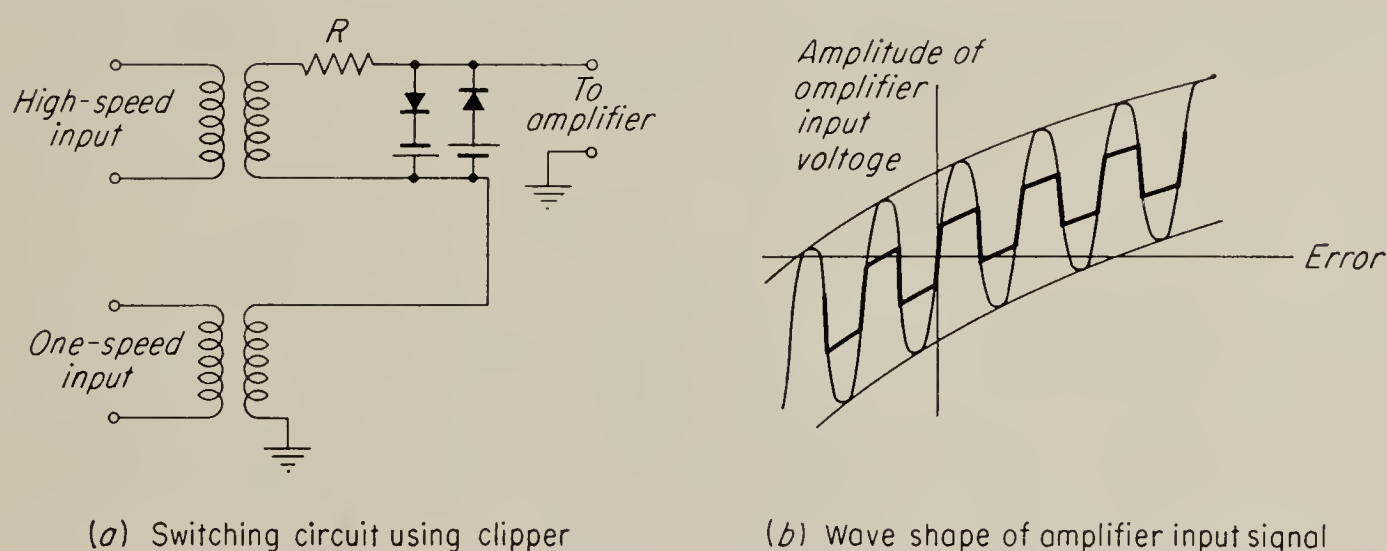


FIG. 5.15. A simple switching circuit for an N -speed synchro.

A switching method not having this advantage but employing a simpler circuit is shown in Fig. 5.15a. The resistor and the rectifiers and batteries constitute a simple clipping circuit, which prevents the signal from the high-speed channel from exceeding a certain level. The one-speed signal is then simply added to the clipped high-speed signal. The form of the composite signal is shown in Fig. 5.15b; note that, if the clipping is sufficiently great, it permits the voltage to go through zero only once in the neighborhood of synchronization. As a result of the sudden change in gain produced by the clipper, a servo using this circuit tends to have relatively poor synchronizing characteristics when *slewing*, i.e., when large and rapid changes of the input are made. Under these conditions the motor may be running at such a high speed as it approaches correspondence that it passes entirely through the region of high-speed control on the first approach, and possibly on several succeeding ones. The system behaves, therefore, as though it were poorly damped, even though the response to small disturbances, which do not require operation beyond the high-speed control zone, is quite well damped. This type of performance is typical of systems in which some component in the loop saturates for relatively small errors, i.e., where the zone of linear operation is small.

Hence there is no objection to using the circuit of Fig. 5.15a if amplifier saturation takes place for smaller error values than those for which clipping occurs.

The two circuits described here are typical but do not exhaust the list by any means. The reader is referred to the literature for a number of other circuits of this sort.¹

An important problem in the design of switching circuits is the determination of the point at which switching is to occur. A criterion for proper operation of the switching circuit may be developed by reference to Fig. 5.16, which shows the variation of error voltage as a function of servo displacement error for the two synchros. Suppose first that the synchros have no static error. This is indicated by the heavy lines in Fig. 5.16. Switching from the one-speed to the N -speed channel must occur for error values on the one-speed synchro between e_1 and $-e_1$, for if switching

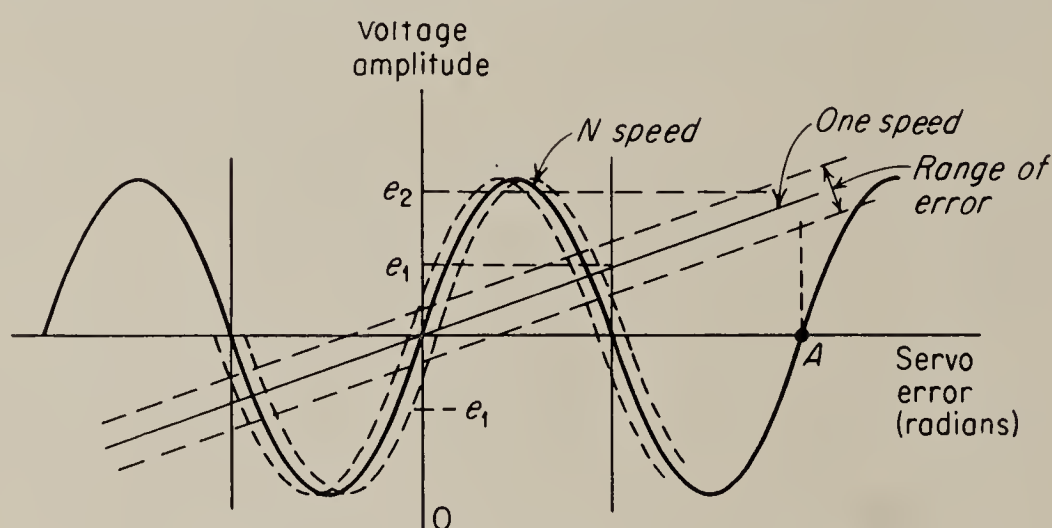


FIG. 5.16. Voltage output of the synchros.

occurs outside this zone, the N -speed synchro will cause the system to move away from the correct correspondence point at O to a false point such as A . The switching circuit normally has hysteresis; i.e., it takes a larger voltage to switch the system to the one-speed channel than it does to switch to the N -speed channel. If the hysteresis is so great that it takes a voltage greater than e_2 to reconnect the one-speed channel, then the system will simply lock in at the false point A . More probably, however, the system switches back to the one-speed channel at a voltage less than e_2 , so that the servo oscillates between the voltages e_1 and e_2 . Neither mode of operation is, of course, desirable.

If e_s is the voltage output of the one-speed control transformer at the point at which the switch connects the system to the N -speed channel, we see from Fig. 5.16 that for proper operation $e_s < k_e \sin(\pi/N)$, or approximately $e_s < k_e \pi/N$, where k_e is the synchro constant. If this inequality

¹ For instance, Ahrendt, "Servomechanism Practice," McGraw-Hill Book Company, Inc., New York, 1954, pp. 59–64; James, Nichols, and Phillips, "Theory of Servomechanisms," Radiation Laboratory Series, vol. 25, McGraw-Hill Book Company, Inc., New York, 1947, p. 84.

is only barely satisfied, then, although the system will lock in at the right place, it may hesitate slightly immediately after the switch has operated, since the voltage from the N -speed synchro is very small. Hence in practice e_s should be about $\frac{1}{2}k_e\pi/N$.

Actually, errors exist in both synchros; in fact it is the existence of these errors that makes the extra complication of the two-speed system necessary. A possible range of one-speed and N -speed error is shown as the dotted lines in Fig. 5.16. Suppose we let ϵ_1 and ϵ_N be the maximum error of the one-speed and N -speed synchro systems, respectively. For identical synchros, ϵ_1 will be approximately equal to ϵ_N . Then, by inspection of Fig. 5.16 and by use of the sort of reasoning used above, we find that for proper operation

$$e_s < k_e \left(\frac{\pi}{N} - |\epsilon_1| - \left| \frac{\epsilon_N}{N} \right| \right) \quad (5.27)$$

or, if $\epsilon_1 = \epsilon_N$,

$$e_s < \frac{k_e}{N} [\pi - (N + 1)|\epsilon_1|] \quad (5.28)$$

The absolute-value bars are used because ϵ_1 and ϵ_N may be either positive or negative. As for the case of zero synchro error, it is desirable to do more than to barely satisfy this inequality.

Since only the magnitude of e_s , not its phase, is usually used to actuate the switching circuit, e_s cannot be negative. Hence inequality (5.28) gives us an upper limit on the gear ratio N that may be employed between the true output and the N -speed selsyn. Specifically

$$N + 1 < \frac{\pi}{|\epsilon_1|} \quad (5.29)$$

Thus for an error of, say, 0.5° (this is rather larger than usually observed, but serves to illustrate the principle), N would have to be less than 360. Actually N seldom exceeds 100 in practice, since the switch cannot be expected to operate always at exactly the same voltage and a safety factor is required. Furthermore e_s cannot actually be zero but must have a finite value. Very large values of N are not justified if the gearing error becomes comparable to the synchro error or if other components of the servo limit the maximum accuracy obtainable. Another factor limiting maximum gear ratios is the fact that the moment of inertia of the control-transformer rotor is reflected by the square of the gear ratio to the motor shaft (see Chap. 8), and with small motors and large ratios this may result in a considerable increase of the effective moment of inertia seen by the motor. Finally, the top design speed of the synchros may limit the maximum gear ratio that should be employed.

One further problem of two-speed synchro operation must be discussed

here: the problem of the 180° ambiguity found in two-speed systems having an even gear ratio. In Fig. 5.17 is shown the variation of output voltage of the two synchros near 0° and 180° for systems having odd and even gear ratios. Considering the odd gear ratio first, we see that the slope of the curves at 180° for both synchros is negative; hence 180° is not a point of stable equilibrium for the servo system. If the servo happened to be energized, with the error exactly 180° , any small disturbance would cause the N -speed voltage to increase and make the system drive away

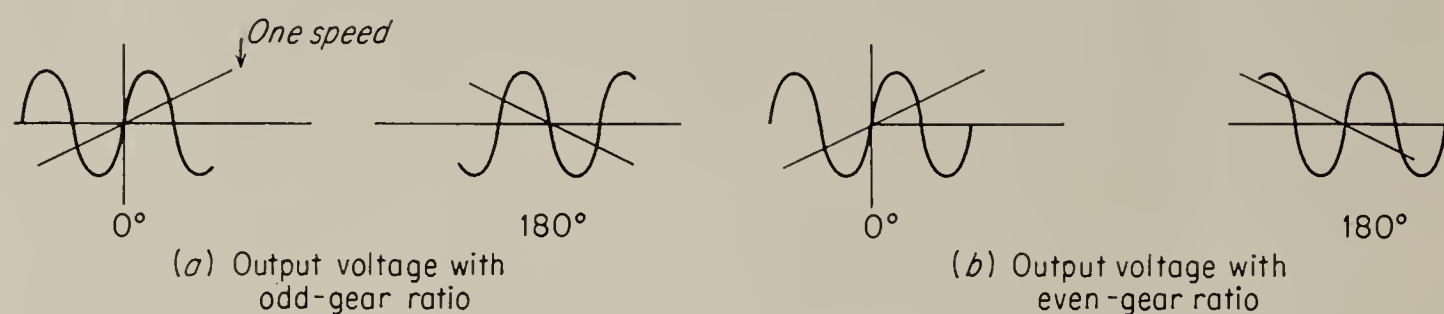


FIG. 5.17. Variation of output voltage on two-speed synchros.

from the 180° position. As the voltage increases, the one-speed system takes over and brings the servo to the proper position at 0° error. For the even ratio, however, the slope of the N -speed curve is positive both at 0° and at 180° . Therefore, since at 180° the one-speed system is switched off, the 180° point is a point of stable equilibrium at which the servo may come to rest.

It is often desirable to use even gear ratios, particularly the ratio 36, since this permits dials calibrated for 360° per revolution and 10° per revolution to be attached directly to the one-speed and 36-speed synchro shafts, respectively. To overcome the difficulty discussed above, a

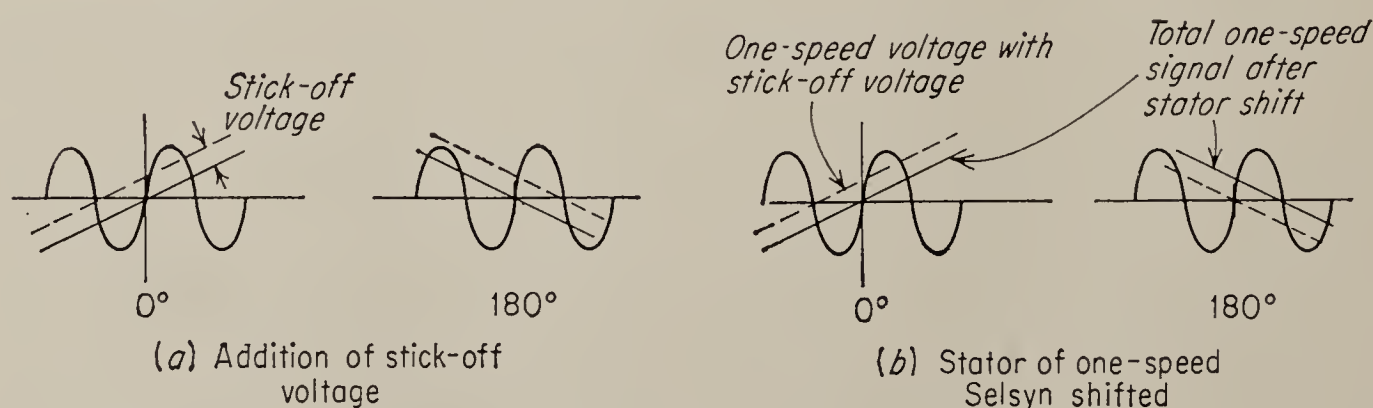


FIG. 5.18. Elimination of false zero at 180° in even-gear-ratio synchro.

constant voltage, the so-called *stick-off voltage*, is added to the output of the one-speed generator. The result of adding this voltage is shown in Fig. 5.18a. The dotted line shows the resultant of adding the stick-off voltage to the one-speed selsyn output. Note that the first effect is that the one-speed and N -speed signals no longer go to zero together at either 0° or 180° . It is necessary that the two signals both be zero when the error is zero, and this may be accomplished by resetting the zero of the one-speed synchro, i.e., by so shifting its stator that the total one-speed

voltage goes to zero at the same point as the N -speed voltage. The effect is shown in Fig. 5.18*b*. Note that with the proper amount of stick-off voltage the signal of both synchros goes to zero near 180° with a negative slope.

It should be noted that the voltage waves of Fig. 5.18 represent the amplitude of a-c signals, with negative amplitude referring to a signal 180° out of phase with one having positive amplitude. Hence the constant stick-off voltage must be a constant a-c voltage, in phase with the synchro output. This voltage is easily added to the one-speed synchro output by means of a small transformer in a circuit, such as the one shown in Fig. 5.19.

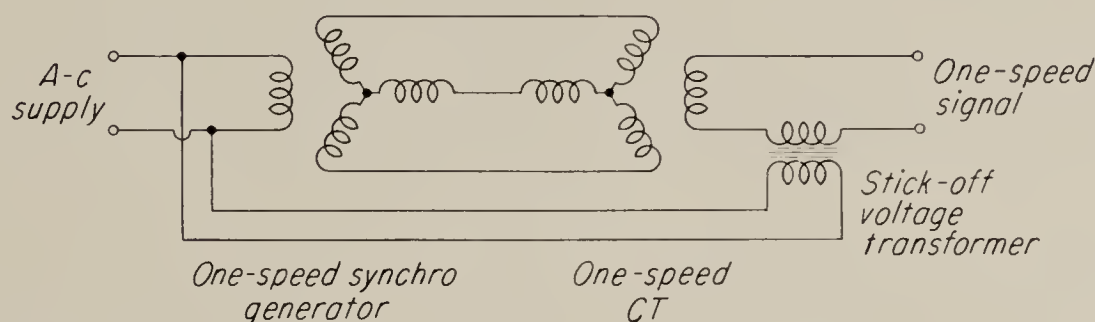


FIG. 5.19. Method of adding stick-off voltage.

5.6. The Differential Generator. When it is necessary to have a servo respond to two or more different shaft-position inputs at various remote locations, differential synchro generators provide a convenient means for adding the additional inputs. Differential synchros are very similar in construction features to synchro generators or control transformers. The primary difference is that the rotor of a differential unit is wound for three phases rather than for single phase, and three slip rings and brushes are required. The rotor is of the slotted-drum type constructed of punched laminations of the sort shown in Fig. 5.3*c*. The windings of both rotor and stator are of the standard, distributed type. Electrically a differential generator, therefore, resembles most closely a wound-rotor induction motor. Despite its constructional similarity to a three-phase machine, it operates on single phase, like the other synchro devices discussed thus far.

A typical circuit showing a differential synchro is shown schematically in Fig. 5.20. Note that the differential synchro is inserted between the generator and control transformer. Its operation is best explained by considering it as a variable mutual inductance inserted between the generator and control transformer. For simplicity, we assume the synchro generator to be perfect, free from errors and unbalances, so that Eqs. (5.13) become

$$\begin{aligned} I_a &= I_m \cos (\theta + 60^\circ) \\ I_b &= I_m \cos (\theta - 60^\circ) \end{aligned} \quad (5.30)$$

In a similar fashion, and by using leg $0'3'$ as the reference axis, we find

$$I'_b = I_m \cos (\theta - \phi - 60^\circ) \quad (5.35)$$

We see that under these simplifying assumptions the form of I'_a and I'_b is the same as that of I_a and I_b , except that the angle $\theta - \phi$ has been substituted for the angle θ . Hence, the effect of the differential generator is to change the control-transformer angle for zero output voltage to $\delta = \theta - \phi$ rather than $\delta = \theta$, the value obtained when the differential generator is not present. Thus the shaft angle of the differential generator is added (or subtracted) from the angle of the input. This may be represented schematically as in Fig. 5.21. It is, of course, possible to use more than one differential generator when more than one additional input must be inserted into the system.

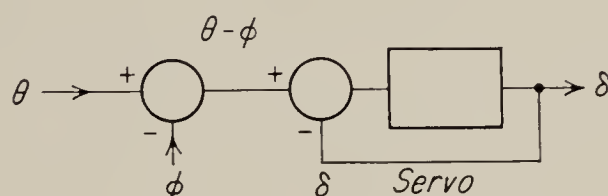


FIG. 5.21. An equivalent block diagram showing the effect of the differential generator.

In practice, differential generators are subject to the same sort of errors as other synchros.¹ Also, the coefficient of coupling between stator and rotor circuits is less than unity, and therefore the differential generator requires a fairly large magnetizing current, which must be supplied by the synchro generator. This current produces heating in the generator, and it places a limit on the number of differential generators that can be excited by a synchro generator of a given size. The magnetizing current may be reduced quite considerably, at least as far as the synchro generator is concerned, by the use of a synchro capacitor. This is a set of three capacitors, usually connected in delta, and “potted” in a single can. They are ordinarily connected across the three lines connecting the synchro generator and differential generator. Synchro capacitors are also sometimes used in simple generator-control-transformer systems, particularly when one generator supplies several control transformers. Since the impedance of the control transformer is usually quite high, a smaller capacitor is required in this application.

5.7. Synchro Repeaters. A synchro repeater (also called synchro motor) is externally and in electrical characteristics identical to a synchro generator. The chief constructional difference is that the rotor of a synchro repeater usually has a vibration damper (see Chap. 8) on it. In typical systems the repeater is connected to the generator as shown in Fig. 5.22. The operation may be explained qualitatively by thinking of the generator as transmitting a magnetic-field pattern to the stator of the

¹ G. H. Weiss and J. H. Rosenbloom, “Linear Lumped Parameter Analysis of Synchros, VIII, Simple Differential Control Systems,” NAVORD Rept. 2569, U.S. Naval Ordnance Laboratory, White Oak, Md., December, 1953.

repeater in a way similar to that explained in connection with the generator-control-transformer system (Sec. 5.2). Since the rotor of the repeater is also excited, it produces another magnetic field, alternating in step with the field established in the stator. Owing to magnetic attraction, the rotor of the repeater will experience a torque tending to line up these two fields, and equilibrium for the system is reached only when the two rotors are parallel. Since the repeater is electrically identical with the generator, the generator rotor experiences the same torque as the rotor of the repeater. The system acts, therefore, as a direct torque transmission. The torque is actually a sinusoidal function of the angular difference between the two rotor positions, but for the small angular differences that are usually of interest the torque may be assumed to be

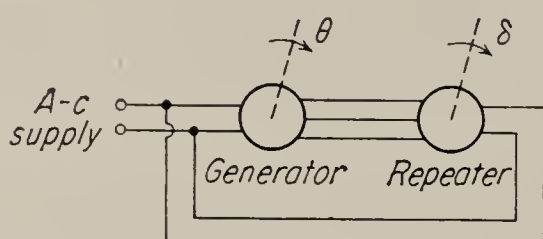


FIG. 5.22. A generator-repeater system.

proportional to the angular difference. The two synchros act, therefore, as though they were coupled together mechanically by a spring. The spring constant of this spring may be referred to as the *torque gradient* of the system and may be computed as a function of the impedances and other parameters of the two machines.¹

Repeaters are commonly used for remote indication of position, in which case their load is only a pointer or a dial. Their accuracy in this application is improved by mounting the rotor on ball bearings and by reducing all sources of friction to an absolute minimum. Since the electrical damping is also inherently very small, a synchro repeater would oscillate for many cycles as a result of any disturbance if it were not equipped with the vibration damper already referred to. Since a machine designed as a repeater is, however, otherwise identical with a generator, it may be used in place of the generator if the presence of the vibration damper is not objectionable.

The generator-repeater combination shown in Fig. 5.22 is only one example of a synchro torque system. Other torque systems are occasionally used. In one of these, two synchro generators and a differential generator (equipped with a vibration damper) are employed. One of the generators feeds the stator and the other the rotor of the differential unit, and the effect is that the rotor of the differential generator tends to take on a position that is equal to the difference between the two synchro-generator rotor positions. The system, therefore, acts like a mechanical differential gear, but since such a system is not very stiff, its characteristics tend to be inferior to those of a mechanical differential.

¹ B. D. Fried and J. H. Rosenbloom, "Linear Lumped Parameter Analysis of Synchros, VII, Simple Torque Systems," NAVORD Rept. 2829, U.S. Naval Ordnance Laboratory, White Oak, Md., July, 1953.

5.8. Variable-reluctance Transducers. All synchros suffer from the disadvantage that an electric connection must be made to the rotor and that brushes and slip rings are required. This disadvantage does not exist in variable-reluctance transducers, and it is therefore possible to reduce friction in them to very low values. On the other hand, variable-reluctance transducers usually can be used only over a limited range of the input position. One of the simplest examples of this type of device is the *E pick-off*, shown schematically in Fig. 5.23. In its most common

form it consists of an E-shaped stator made by stacking a number of E-shaped iron laminations. A coil is wound on each one of the three legs, and the two outer coils are connected in opposition. The center coil is normally excited by a constant a-c voltage. When the

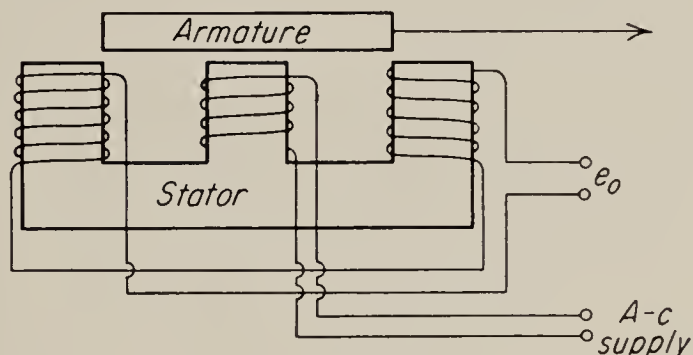


FIG. 5.23. An E pick-off.

movable armature, also made of iron laminations, is symmetrically located with respect to the two outer legs, the mutual inductance between each of the outer legs and the central leg is the same, and since the outer coils are wound in opposition, no output voltage results. This position of the armature is referred to as the *null position*. If the armature is displaced to the right from the null position, the voltage induced in the right-hand coil is larger than that induced in the left-hand coil; hence an output voltage is produced that is ideally in phase with the applied voltage. Motion of the armature to the left produces an output that is 180° out of phase with the voltage produced by motion to the right. Within limits the output voltage is approximately proportional to the displacement; hence we have a device that modulates the a-c supply voltage in accordance with the displacement of the armature. The modulation provided is of the suppressed-carrier type. If the armature position is made to oscillate sinusoidally about the null position, an output-voltage waveform of the sort shown in Fig. 5.5 results.

In practice, E pick-offs depart in a number of ways from the ideal behavior described above. For example, they usually produce a fairly large quadrature voltage at the point where the output voltage is theoretically zero. Qualitatively, the reasons for the existence of this voltage are similar to those already discussed in relation to synchros (Sec. 5.3). The quadrature voltage can often be reduced markedly by connecting a small variable capacitance between one of the leads to the center coil and one of the output leads (see Fig. 5.24). The purpose of the isolating transformer shown in Fig. 5.24 is to permit connecting the capacitor to either side of the input coil. Occasionally, no adjustment of the capaci-

tor completely removes the null voltage. In this case a high-resistance potentiometer may be connected across the center coil, with the tap going to the ungrounded lead of the output. This provides a second adjustment, and by adjusting both the potentiometer and the capacitor simultaneously, the voltage existing at the null position can be reduced to a residue consisting only of harmonics of the supply voltage. If these are still objectionable (as they sometimes are in very high gain systems), the output signal may be passed through a bandpass filter adjusted to pass only the fundamental frequency. E pick-offs are also subject to static and velocity errors just as synchros are. However, since E pick-offs are usually used singly, the type of error in synchros arising from the fact that these units are used in pairs is of no consequence. Also, since E pick-offs have only a limited range of displacements, the velocity error is

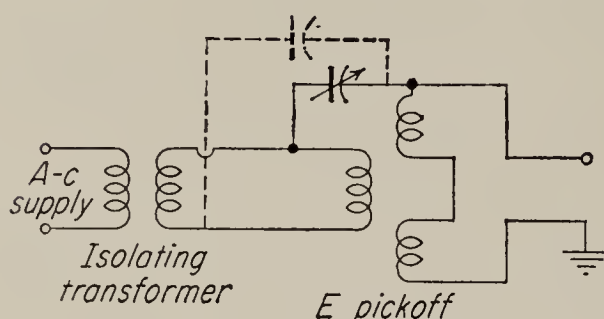


FIG. 5.24. Schematic of E pick-off showing capacitor connected to reduce quadrature voltage.

unimportant. The primary accuracy consideration of importance with E pick-offs is, therefore, linearity. Linearity depends on the way in which the reluctance changes between the center leg and the outer legs, and the most linear system will result if the reluctance decreases as much on one side as it increases on the other. This can be accomplished by making the length of the armature such that at the null position it covers approximately half of each of the outer legs. It is also important that the reluctance of the iron, which is a nonlinear function of flux density, be negligibly small compared to the air-gap reluctance; hence a relatively long air gap is desirable if linearity is important. Greater sensitivity at the expense of reduced linearity can, however, be obtained by reducing the air gap to a minimum and by shortening the armature so that at the null position it just barely covers the two outer legs (as in Fig. 5.23).

Although the E pick-off described here is designed to convert linear motion into an electric signal, it can be used also for rotary motion by shaping the moving armature and the E-shaped stator as shown in Fig. 5.25. It is clear, of course, that only a limited range of angular displacement can be handled by this pickup, but in many null-type systems, where the stator and armature are moved directly by the input and output, respectively, and where relative displacement between stator and armature represents the error of the feedback loop, this is no disadvantage.

A somewhat different form of variable-reluctance transducer that is very closely related to the E pick-off is shown in cross section in Fig. 5.26. Three coils are used, just as in the E pick-off, but they are wound on a fiber or other nonmagnetic and nonconducting core. The two outer coils

are connected in opposition. A small rod, carrying a short iron section, passes through the center of the coils as shown in the figure and performs the same function as the armature in the E pick-off. Normally the center coil is connected to the supply and acts as the primary, and the signal is taken from the two outer coils. The operation is exactly the same as that shown for the E pick-off. This transducer can be kept very small in

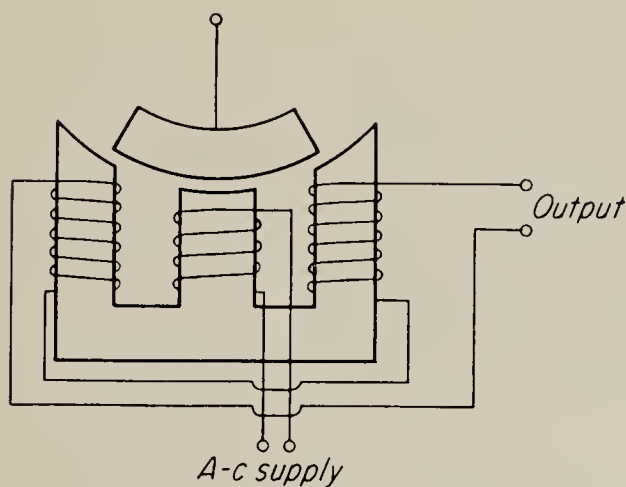


FIG. 5.25. A schematic of an E pick-off designed for rotary motion.

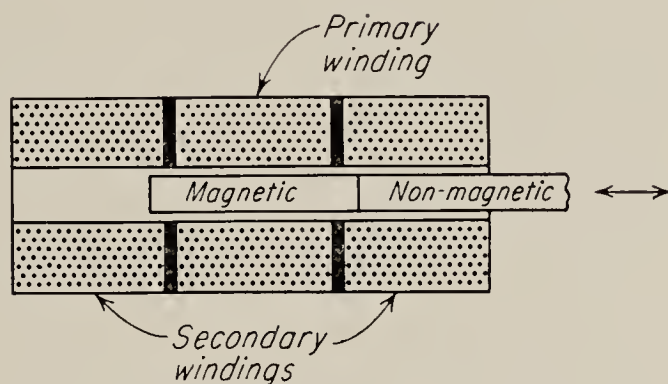


FIG. 5.26. Schematic of a variable-reluctance transformer.

size and is, therefore, often used in miniature equipment. It is commonly used in accelerometers and pressure pickup devices as the element that converts the minute displacements obtained in these devices into an electric signal.

5.9. The Microsyn. Another device very closely related to the E pick-off is the *microsyn*, shown schematically in Fig. 5.27. Both the stator and rotor are made of iron laminations, the rotor being specially shaped to vary the reluctance of the stator magnetic circuit. The stator has four poles, and each pole carries two coils, a primary and a secondary coil. All the primaries are connected in series and are so wound on the poles that, for a particular direction of current in the winding, the flux vector is inward on poles 1 and 2 and outward on poles 3 and 4. The secondaries are also all in series but are connected in such a way that the voltages induced in the windings on poles 1 and 3 oppose the voltages induced on poles 2 and 4. Hence, when the rotor is in the neutral position as shown in the figure, the flux in all four poles is the same, and no output voltage is produced. However, motion of the rotor away from neutral results in a voltage output, which over a limited range of rotation (about 7°) can be

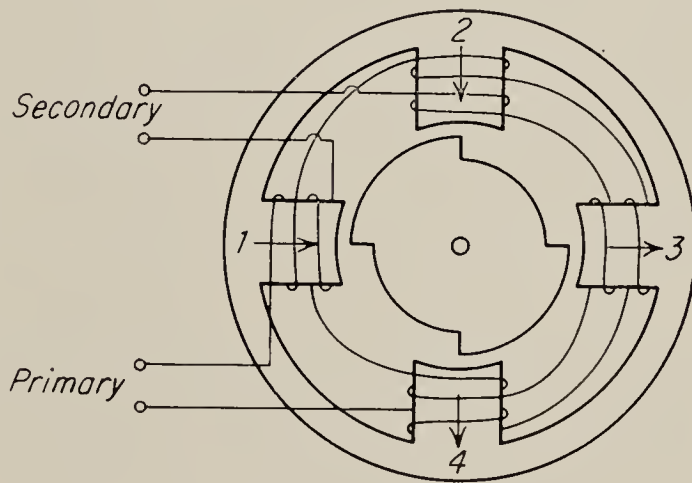


FIG. 5.27. The microsyn.

made quite accurately proportional to the angle. The microsyn is used normally only for limited angular rotations.

There are a number of other devices of the variable-reluctance type such as the *telegon*, the *magnesyn*, and others. The reader is referred to the literature¹ for descriptions of these transducers.

PROBLEMS

5.1. The inductive reactance of each leg of the stator winding of a type 1 CT control transformer is 1,000 ohms and that of the type 1 G generator used to excite it is 100 ohms. Nominally the ratio of resistance to reactance in both machines is 1:4, but as a result of a bad connection in one stator lead, the resistance ratio in this one leg is increased to 1:3.9. Find the maximum quadrature voltage from the control transformer if the maximum inphase output voltage is 57 volts rms.

5.2. Determine the maximum second-harmonic error in a synchro transmission in which the ratio of $I_{am}/I_{bm} = 0.99$. Assume that all other errors, including the quadrature error, are zero.

5.3. Find the maximum value of sixth-harmonic error in a synchro transmission in which the fifth space harmonic is 20 per cent of the fundamental and the seventh 10 per cent of the fundamental. Assume all other errors negligible.

5.4. A servo system using a two-speed synchro transmission is supposed to have an effective velocity-error constant of at least 1,000. The ratio of inductive reactance to resistance in the control transformers is 4. Assuming that the servo gain constant cannot exceed 5,000, find the maximum gear ratio permissible between the output shaft and the high-speed synchro.

5.5. A servo system uses a two-speed synchro transmission with a 36:1 ratio between high- and low-speed synchros. The error in both synchros may be as large as 0.3° ; the synchro constant for both is 1 volt/deg. The switching circuit used to switch the system between the high- and low-speed synchros has the form shown in Fig. 5.14. The neon tubes fire when an rms voltage of 60 volts is applied to them; they go out when the voltage is 40 volts. Assume that the transformers T_1 and T_3 have unity turns ratio; find the extreme values of the combined amplifier gain and transformer T_2 turns ratio that will permit reliable switching.

5.6. What is the value of the stick-off voltage required to prevent ambiguity in a system using a two-speed synchro system with 36:1 gear ratio? The synchro constant for both synchros is 1 volt/deg. Neglect synchro error.

¹ See particularly Greenwood, Holdam, and MacRae, "Electronic Instruments," Radiation Laboratory Series, vol. 21, McGraw-Hill Book Company, Inc., New York, 1948, pp. 362-370. Also Batcher and Moulic, "Electronic Control Handbook," Caldwell Clements, Inc., New York, 1947.

CHAPTER 6

DEMODULATORS AND MODULATORS

6.1. Fundamental Concepts. The output signal delivered by a synchro control transformer or by any of the electromagnetic transducers described in the previous chapter is in the form of a relatively high frequency carrier modulated by signal information that is normally confined to a much lower frequency spectrum. Although there are some servo power units, such as the a-c motors considered in the next chapter, that can utilize this modulated signal directly, it is usually necessary to recover the low-frequency envelope by use of a *demodulator* or *discriminator*. Since the output signal delivered by most of the transducers used in servomechanisms is of the suppressed-carrier, amplitude-modulated type, demodulators used with these transducers must be *phase-sensitive*; i.e., they should put out a positive voltage when the input is in phase with a given reference and a negative voltage when the input is 180° out of phase.

A simple circuit to accomplish this type of demodulation is shown in Fig. 6.1. This circuit is essentially a push-pull amplifier of standard

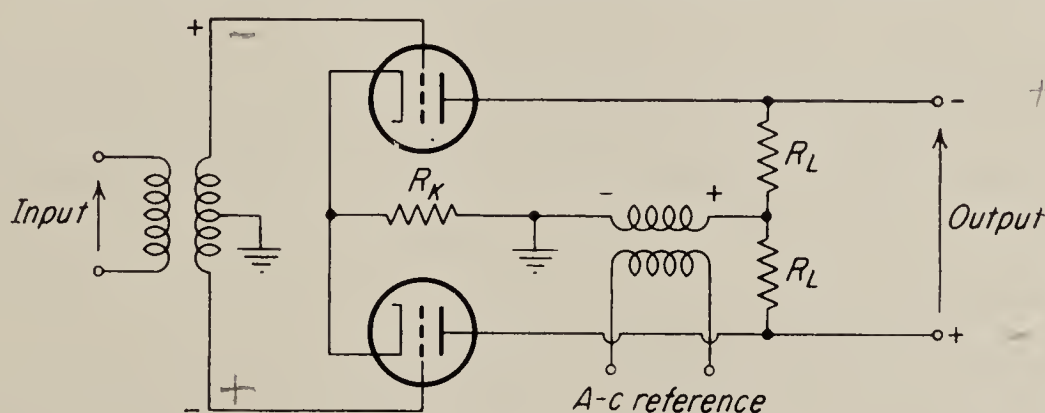


FIG. 6.1. Half-wave triode demodulator.

design with plate load R_L and cathode-biasing resistor R_K . The feature that makes the circuit act as a phase-sensitive demodulator is the a-c plate-supply voltage. This supply is referred to as the *reference voltage*, and it must be of the same frequency as the input signal. As a result of the a-c supply, the amplifier is cut off every other half cycle, operating only during the half cycle in which the plate voltage is positive. If the phase of the input signal is such that during the on period of the amplifier the upper grid in Fig. 6.1 is positive and the lower one negative, more current flows in the upper triode than in the lower, and hence the

plate of the upper triode is more negative than that of the lower. If the phase of the input is reversed, the polarity of the output reverses. Typical output waveshapes obtained from the circuit for various phase relations between input and reference are shown in Fig. 6.2. In drawing these figures, it is assumed that the gain of the tubes is independent of the magnitude of the plate voltage; this is a good assumption during most of the time that the voltage is positive. Figure 6.2c shows the effect of an input signal 90° out of phase with the reference; note that the average, or d-c, output is zero. The discriminator is therefore seen to reject quadrature signals. In general, if the input is less than 90° out of phase with the reference, the d-c component of the output is reduced from the value it would have if the signal and reference were exactly in phase. A typical output waveform is shown in Fig. 6.2d.

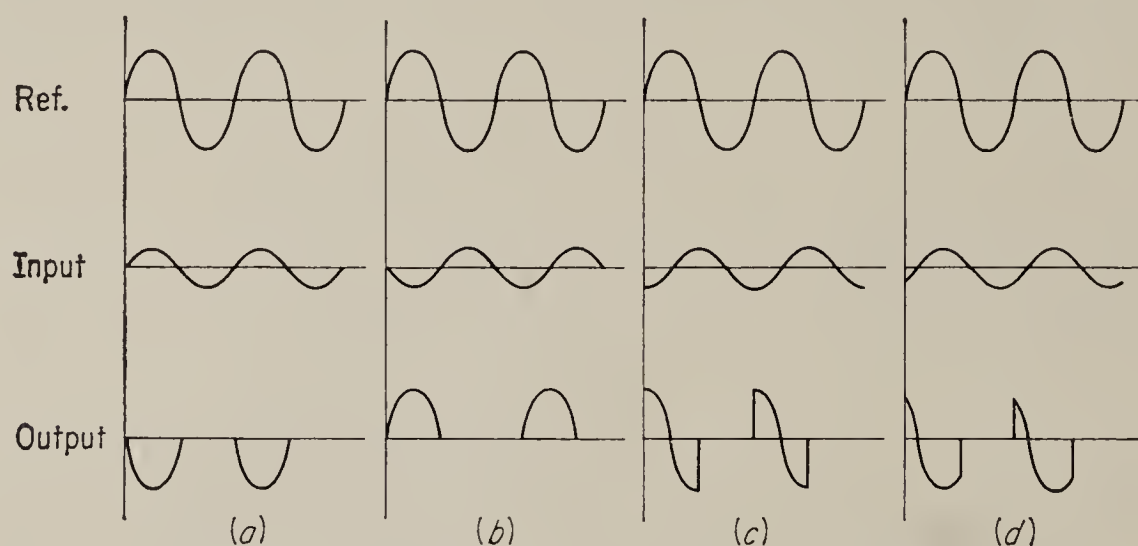


FIG. 6.2. Demodulator output waveshapes.

In all cases shown, the output contains a large a-c ripple component. This ripple is normally highly objectionable because it results in saturation, heating, and general malfunctioning of amplifiers and power devices following the discriminator. Hence low-pass filters are usually used in the discriminator output. For most applications these are simple RC networks of the type described in Chap. 1.

The output waveforms shown in Fig. 6.2a and 6.2b are those obtained from a standard half-wave rectifier. Hence the discriminator circuit shown in Fig. 6.1 is referred to as a *half-wave discriminator*. Some of its other features are that it provides gain between input and output and that it delivers a push-pull output signal. If the two halves of the circuit are exactly identical, zero input results in both zero d-c and zero ripple output. In practice, however, perfect balance of the circuit is difficult to achieve, and there is usually some output for zero input. Since the circuit unbalance may vary as a result of temperature and supply-voltage changes, the output produced by the unbalance may also vary; i.e., the circuit may drift. This is not unreasonable since the circuit is in part a

d-c amplifier, but it should be noted that since it is a push-pull circuit, the drift will be quite small.

6.2. Analysis of the Half-wave Discriminator.¹ In analyzing the operation of the discriminator circuit considered in the preceding paragraph, it is convenient to assume that the output voltage delivered during the on period of the amplifier is equal only to the amplified input signal and is independent of the plate voltage. While this is not quite true, it is approximately correct during the major part of the on period if the plate characteristics of the tube in question are fairly linear, that is, if r_p and μ are approximately independent of the operating point. Also, it will be found that the exact form of the output signal will affect the analysis only in detail, not in principle.

If we let the gain during the on period be G , then under the above assumption the operation of the circuit is simply explained by considering it to multiply the input signal by the square wave shown in Fig. 6.3. This is a square wave having an amplitude of $G/2$ plus a d-c component of $G/2$. It may be expanded into the Fourier series:

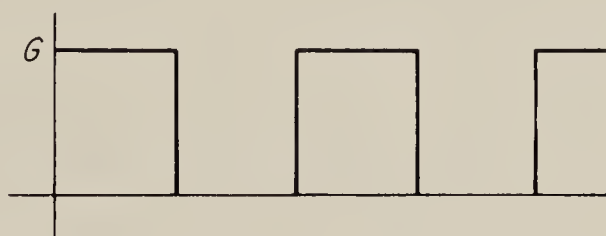


FIG. 6.3. Approximate demodulating function.

$$f_d(t) = \frac{G}{2} + \frac{2G}{\pi} (\sin \omega t + \frac{1}{3} \sin 3\omega t + \frac{1}{5} \sin 5\omega t + \dots) \quad (6.1)$$

where ω is the reference or carrier frequency in radians per second.

The input signal applied to the discriminator is assumed to be suppressed-carrier amplitude-modulated, and for the moment we assume that the carrier is a sinusoidal function of time. If the modulating function is $F(t)$, then the input signal to the discriminator is given by

$$v_i(t) = F(t) \sin (\omega t - \beta) \quad (6.2)$$

where ω is the carrier frequency and β is the phase shift that might exist between the carrier and the reference. The output of the discriminator is the product of the input and the demodulating gain function of Eq. (6.1):

$$v_o(t) = GF(t) \sin (\omega t - \beta) \left[\frac{1}{2} + \frac{2}{\pi} \left(\sin \omega t + \frac{1}{3} \sin 3\omega t + \frac{1}{5} \sin 5\omega t + \dots \right) \right] \quad (6.3)$$

In expanding this expression, it is convenient to expand first the factor $\sin (\omega t - \beta)$ into $\sin \omega t \cos \beta - \cos \omega t \sin \beta$. Multiplying this into the series and simplifying results in

¹ J. L. Bower, "Discriminators and Modulators for Servomechanisms," unpublished report, 1945.

$$\begin{aligned}
v_o(t) = GF(t) \left\{ \cos \beta \left[\frac{1}{\pi} + \frac{1}{2} \sin \omega t - \frac{2}{\pi} \left(\frac{1}{3} \cos 2\omega t + \frac{1}{3 \cdot 5} \cos 4\omega t \right. \right. \right. \\
\left. \left. \left. + \frac{1}{5 \cdot 7} \cos 6\omega t + \dots \right) \right] \right. \\
\left. - \sin \beta \left[\frac{1}{2} \cos \omega t + \frac{4}{\pi} \left(\frac{1}{3} \sin 2\omega t + \frac{2}{3 \cdot 5} \sin 4\omega t + \frac{3}{5 \cdot 7} \sin 6\omega t + \dots \right) \right] \right\}
\end{aligned} \tag{6.4}$$

The useful part of the output is provided by the first term in the first bracket.

$$\text{Useful } v_o(t) = \frac{G}{\pi} (\cos \beta) F(t) \tag{6.5}$$

All other terms represent ripple. We note that the original modulation $F(t)$ has emerged from the discriminator multiplied only by the constant factor $(G/\pi) \cos \beta$, without any phase lags or amplitude changes depending on frequency. We conclude from this that the processes of modulation and demodulation do not inherently introduce any phase lag (or lead) into the signal channel and that the phase lag normally found is entirely due to the ripple filters. As mentioned previously, these are almost always required to eliminate the objectionable effects of the ripple on the circuits following the discriminator.

Equation (6.5) shows that the gain between input and useful output is proportional to the cosine of the phase shift between the carrier and the reference. The desirability of keeping this phase angle small if maximum gain is to be secured is therefore evident. However it is also clear that phase shifts of the order of less than 10° are not of any great consequence. Phase shift also has some effect on the magnitude of the ripple. This can be appreciated qualitatively by inspection of Eq. (6.4), since the second bracket, which contributes a large ripple component, is proportional to the sine of the phase angle and therefore vanishes if the carrier and reference are in phase. A somewhat more quantitative idea of the effect of phase angle on output ripple is obtained by computing the magnitude of the harmonics from Eq. (6.4) as a function of β . The results of such a computation are given in Table 6.1 for the case of $GF(t) = 1$. The table

TABLE 6.1. THE EFFECT OF PHASE ANGLE ON THE MAGNITUDE OF OUTPUT RIPPLE

<i>Harmonic</i>	<i>Magnitude</i>
0	$0.318 \cos \beta$
1	0.5
2	$0.21 \sqrt{1 + 3 \sin^2 \beta}$
4	$0.043 \sqrt{1 + 15 \sin^2 \beta}$
6	$0.018 \sqrt{1 + 35 \sin^2 \beta}$
8	$0.01 \sqrt{1 + 63 \sin^2 \beta}$

shows clearly that the ripple voltage increases with β , but since the largest ripple-frequency component, the fundamental, is independent of β , one would not expect the rms value of the ripple to increase very much as β changes from 0 to 90° . It is instructive in this connection to compute the rms value of the ripple for these two extreme values of β . Using only the harmonics listed in Table 6.1, we find approximately that for $\beta = 0^\circ$ the rms value of the ripple is 0.544, while for $\beta = 90^\circ$ the ripple increases to 0.70. Thus it is clear that, although phase shift does increase the ripple somewhat, the effect is not of any great importance in a half-wave discriminator.

In order to find the maximum frequency of the input modulation $F(t)$ that can be handled without ambiguity by a half-wave discriminator, we let

$$F(t) = V \sin \omega_s t \quad (6.6)$$

Substitution of this expression into (6.4) results, after some simplification, in

$$v_o(t) = GV \left\{ \cos \beta \left[\frac{1}{\pi} \sin \omega_s t + \frac{1}{4} \cos (\omega - \omega_s)t - \frac{1}{4} \cos (\omega + \omega_s)t - \cdots \right] \right. \\ \left. - \sin \beta \left[\frac{1}{4} \sin (\omega - \omega_s)t + \frac{1}{4} \sin (\omega + \omega_s)t + \cdots \right] \right\} \quad (6.7)$$

The term yielding the useful part of the output is still the first term of the first bracket, $GV (\cos \beta)(1/\pi) \sin \omega_s t$, with all else constituting ripple. Thus the lowest harmonic of the ripple is shifted downward by the signal frequency and becomes $\omega - \omega_s$. It is clear that, when

$$\begin{aligned} \omega - \omega_s &= \omega_s \\ \text{or} \quad \omega_s &= \frac{1}{2}\omega \end{aligned} \quad (6.8)$$

the signal frequency and the frequency of the lowest harmonic of the ripple are the same, and it is no longer possible to distinguish between signal and ripple. Hence we have the important and quite universal result that the absolute upper frequency limit of a half-wave discriminator is equal to one-half the carrier frequency. Practical reasons such as the difficulty of constructing a ripple filter with a very sharp cutoff characteristic usually restrict the maximum permissible signal frequencies to values considerably below this absolute limit. Thus a commonly observed rule of thumb applied to the design of feedback systems is that the loop-gain crossover frequency should be no more than one-tenth of the carrier frequency if a half-wave discriminator is used.

Equation (6.7) contains additional information that is useful in the design of ripple filters. Usually these are simple, low-pass RC networks

of the type described in Chap. 1. However it is sometimes found that a simple filter of this sort designed to attenuate the ripple to a satisfactory degree also results in excessive attenuation and phase shift in the signal spectrum. In such cases, advantage may be taken of the fact that the ripple is concentrated in frequency bands near the fundamental, second, fourth, etc., harmonics of the carrier. This fact permits the use of a series of sharply tuned band-rejection filters, such as the bridged-T or twin-T circuits discussed in Chap. 1, which produce much less phase lag and attenuation at lower frequencies than the simple low-pass filter does. In fact, it can be shown that the phase shift produced by these filters can be made to approach zero as the rejection band is reduced to zero width.¹ It would appear, therefore, that this might be the ideal adjustment. However, if more terms of the series indicated in Eq. (6.7) are written, it will be found that the frequency components found in the ripple are not simply the fundamental, second, fourth, etc., harmonics of the carrier, but are actually $\omega \pm \omega_s$, $2\omega \pm \omega_s$, $4\omega \pm \omega_s$, etc. Hence, if the input $F(t)$ is confined to a spectrum ranging between d-c and some maximum frequency ω_{sm} , the ripple will be found to be concentrated in bands of width $2\omega_{sm}$ centered at the fundamental, second harmonic, etc. The ripple filters should attenuate all frequencies in this band, and hence they cannot be of zero width. We conclude, therefore, that in practice there is a certain minimum amount of phase lag contributed by a frequency discriminator even if the most sophisticated type of ripple filter is employed.

According to Table 6.1 the fundamental and second-harmonic frequency components contribute the major share of the ripple. Hence, if a band-rejection type of filter is used to minimize the low-frequency phase lag, it is usually sufficient to attenuate only the fundamental and the second harmonic by this type of network, with a simple low-pass structure to attenuate all the higher harmonics. Since this low-pass filter must attenuate only the fourth and higher harmonic components, its passband may be relatively wide. Thus the phase shift contributed by this part of the filter at signal frequencies can also be reduced to a minimum. In this way it is possible to obtain good attenuation of all the ripple with relatively small effect on the signal band.

6.3. Full-wave-discriminator Circuits. The analysis of the half-wave discriminator just concluded has shown that its upper frequency limit is equal to one-half the carrier frequency and that the lowest harmonic of the ripple frequency is the fundamental of the carrier. Intuitively, it appears that the full-wave discriminator, for which there is an output every half cycle of the carrier, should be able to improve on this performance, and it will be shown presently that this is indeed the case. We con-

¹ Valley and Wallman, "Vacuum Tube Amplifiers," Radiation Laboratory Series, vol. 18, McGraw-Hill Book Company, Inc., New York, 1948.

sider first a typical circuit, operating on the same principle as the one shown in Fig. 6.1, and discuss its operation qualitatively.

The circuit of a triode full-wave discriminator is shown in Fig. 6.4. The dots shown on the transformers are the conventional polarity markings; hence, when the upper two triodes conduct, the lower two are disabled by negative plate supply, and vice versa. Suppose now that the phase of the input is such that the grid of the uppermost tube is positive when the upper two triodes conduct. Then the output signal e_o is positive in the direction shown by the arrow. During the next half cycle the lower two triodes conduct, and the connection of the input transformer is such that the lowermost grid is now negative. Hence the output signal is again

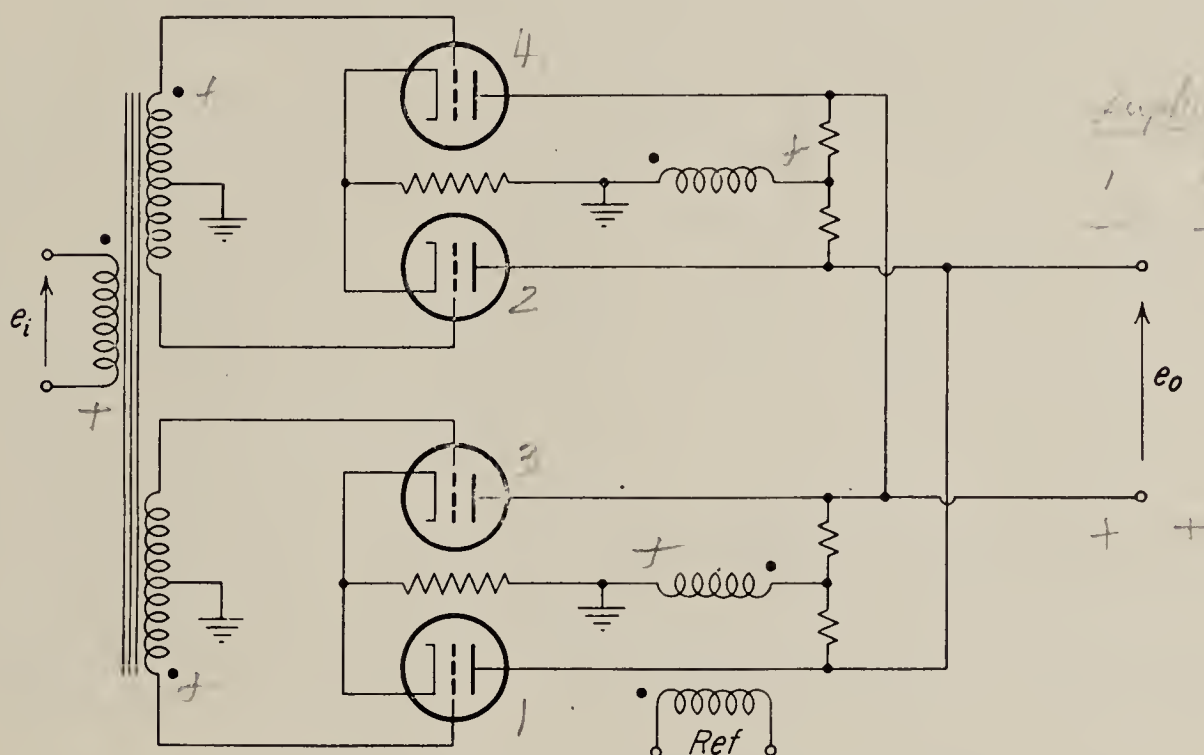


FIG. 6.4. Full-wave discriminator.

positive in the direction of the arrow. Thus for each cycle of the reference we get two pulses of the output, or a full-wave signal. Had the phase of the input relative to the reference been inverted, so that the uppermost grid was negative during the conduction period of the upper two tubes, we would have found the output to consist of negative pulses. If the phase of the input were such that during the conducting period of the upper triodes the uppermost grid were first positive and then negative, the output would consist of a pulse first positive and then negative. By this sort of reasoning one can obtain the output waveshapes shown in Fig. 6.5. We note again that a signal 90° out of phase with the reference will result in zero output, and if the phase shift between signal and reference is somewhere between 0 and 90° , a reduced d-c output results.

In analyzing this circuit we make the same assumption we made in the analysis of the half-wave discriminator, namely, that the gain of the circuit is a constant during the conduction period of either set of triodes. Also, as has already been noted, the operation of the circuit is such that,

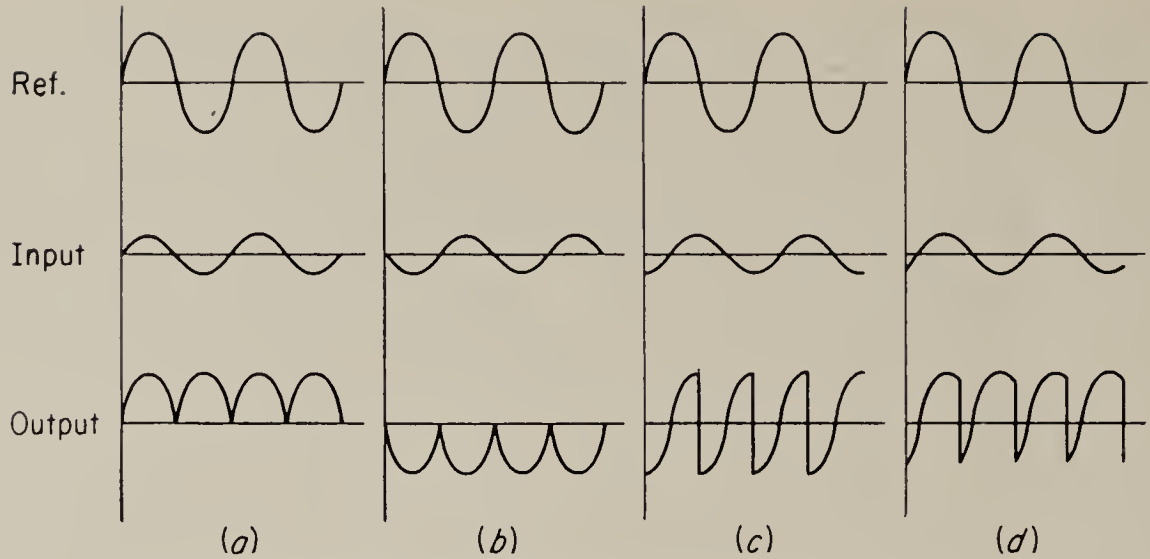


FIG. 6.5. Output waveforms of full-wave discriminator.

if during one half cycle of the reference the output is positive for positive input, then for the next half cycle the output is positive for negative input.

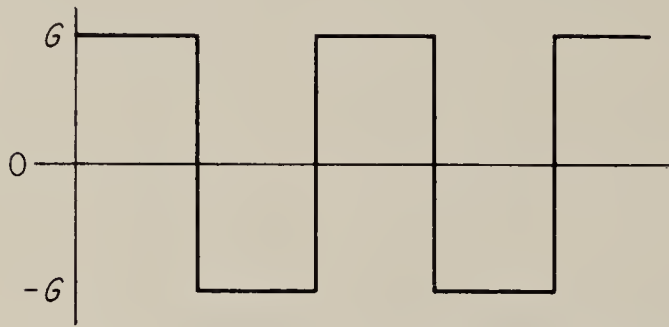


FIG. 6.6. Approximate demodulating function of full-wave discriminator.

Hence, the circuit can be assumed to multiply the input by the square wave shown in Fig. 6.6; this differs from the square wave considered in connection with the half-wave discriminator primarily in that its d-c level is zero. The magnitude is arbitrarily taken as G , the gain of the amplifier in one of the switching modes.

The discriminator gain function may be expanded in the Fourier series:

$$f_d(t) = \frac{4G}{\pi} (\sin \omega t + \frac{1}{3} \sin 3\omega t + \frac{1}{5} \sin 5\omega t + \cdots) \quad (6.9)$$

If the input to the discriminator is again

$$v_i(t) = F(t) \sin (\omega t - \beta) \quad (6.2)$$

then the output given by the product of (6.2) and (6.9) may be expanded into the form

$$\begin{aligned} v_o(t) = GF(t) & \left\{ \cos \beta \left[\frac{2}{\pi} - \frac{4}{\pi} \left(\frac{1}{3} \cos 2\omega t + \frac{1}{3 \cdot 5} \cos 4\omega t \right. \right. \right. \\ & \left. \left. \left. + \frac{1}{5 \cdot 7} \cos 6\omega t + \cdots \right) \right] \right. \\ & \left. - (\sin \beta) \frac{8}{\pi} \left(\frac{1}{3} \sin 2\omega t + \frac{2}{3 \cdot 5} \sin 4\omega t + \frac{3}{5 \cdot 7} \sin 6\omega t + \cdots \right) \right\} \quad (6.10) \end{aligned}$$

The useful part of the output is $F(t)$ multiplied by the time-invariant part

of the series,

$$\text{Useful } v_o(t) = \frac{2}{\pi} G(\cos \beta) F(t) \quad (6.11)$$

with all other terms contributing only to the ripple. Here again it may be seen that the input modulation $F(t)$ emerges unchanged by the processes of modulation and demodulation. Also the effect of β on the gain is identical with that found already for the half-wave discriminator. In fact, comparison of Eqs. (6.4) and (6.10) shows that, except for a factor of 2, which can be absorbed into the gain factor G , the only difference between the outputs of the full-wave and half-wave circuits is that in the former, the fundamental frequency component of the ripple is absent. Hence the table of harmonic magnitudes constructed for the half-wave discriminator (Table 6.1) applies also to the full-wave discriminator if we set $GF(t) = 1/2$ and let the first harmonic be zero.

The table may again be used to find the approximate rms value of the ripple voltage and to estimate the effect of β on it. Using only the harmonics up to and including the eighth, we find that for $\beta = 0^\circ$ the rms value of the ripple is 0.216, while for $\beta = 90^\circ$ it goes up to 0.473. (Note that these values are computed for $GF(t) = 1/2$ in order to effect a direct comparison with the half-wave circuit.) It is clear that for $\beta = 0$ the ratio of rms ripple voltage to useful output is less than half as large in the full-wave as it is in the half-wave circuit. This is due to the absence of the fundamental, which is responsible for the major share of ripple voltage in the half-wave discriminator. It is also apparent that the effect of phase shift between signal carrier and reference is more serious in increasing the ripple, and it is therefore somewhat more important here to keep this phase angle low.

We note that the amount of filtering required to reduce the ripple output following a full-wave discriminator to a permissible amount is much less than that needed to produce an equal attenuation of the ripple produced by a half-wave circuit. This is due both to the smaller rms value of the ripple and to the fact that the lowest ripple harmonic is the second harmonic rather than the fundamental of the carrier frequency. Specifically, let us assume that a double-section RC filter, producing an attenuation proportional to the square of the frequency, is employed as a filter. Then, for the same reduction in ripple relative to d-c output, the filter used with the full-wave circuit may have about three times the passband (as defined by the reciprocal of the RC time constant of the two filter sections) of the filter used with the half-wave circuit. The superiority of the full-wave discriminator in causing reduced attenuation and phase lag of desired signal information is, therefore, evident.

In order to find the maximum frequency of the input modulation $F(t)$

that can be handled without ambiguity by a full-wave discriminator, we again let

$$F(t) = V \sin \omega_s t \quad (6.12)$$

Then Eq. (6.10) becomes

$$v_o(t) = GV \left\{ \cos \beta \left[\frac{2}{\pi} \sin \omega_s t + \frac{2}{3\pi} \sin (2\omega - \omega_s)t - \frac{2}{3\pi} \sin (2\omega + \omega_s)t + \cdots \right] - (\sin \beta) \frac{4}{\pi} \left[\frac{1}{3} \cos (2\omega - \omega_s)t - \frac{1}{3} \cos (2\omega + \omega_s)t + \cdots \right] \right\} \quad (6.13)$$

As before, the first term of the series, $GV (\cos \beta)(2/\pi) \sin \omega_s t$, represents the useful signal, with all other factors being ripple, and it can be seen that the lowest harmonic of the ripple is $2\omega - \omega_s$. The upper limit on ω_s is reached when this lowest ripple frequency coincides with the signal frequency, i.e., when

$$\omega_s = 2\omega - \omega_s$$

or

$$\omega_s = \omega \quad (6.14)$$

Thus the absolute upper limit is equal to the carrier frequency, a limit twice that of the half-wave discriminator. As was mentioned in connection with the half-wave circuit, the actual useful upper limit is normally considerably less than the theoretical maximum because of practical difficulties of filter design. However, if these difficulties are assumed to be identical in the two cases, the fact that the practical upper frequency limit of the full-wave discriminator is twice as high as that of the half-wave discriminator still holds true.

It should be pointed out that the result concerning the maximum theoretical frequency usable with a full-wave discriminator is not universally true but depends on the fact that the input signal was assumed to have a sinusoidal carrier. If the carrier were an accurate square wave, identical with the one assumed for the discriminator demodulating function (Fig. 6.6), the processes of modulation and demodulation would have subjected the signal to two successive multiplications by a square wave. This, at least theoretically, would be equivalent to multiplication by a constant, so that there would be no upper frequency limit and, incidentally, no ripple in the output. While this result is primarily of theoretical interest, owing to the difficulty of generating the required accurate square waves, it does point out the desirability of square-wave modulation and demodulation in general.

As discussed in connection with the half-wave discriminator, it is possible to take advantage of the fact that the ripple is concentrated in relatively narrow bands around the harmonics of the carrier to design an improved type of ripple filter incorporating band-rejection filter sections. The use of this type of filter is of considerably greater practical importance

with the full-wave than with the half-wave discriminator. For the half-wave discriminator the extra complication involved hardly seems justified, since the discriminator itself is not optimum and there is little point in trying to improve an inferior circuit when a better circuit is available. This argument does not apply, however, to the full-wave discriminator, for it represents the best that can be achieved without increasing the circuit complexity manifold. Since the lowest harmonic of the ripple is the second and since the next one (the fourth) is almost five times smaller (see Table 6.1), a relatively large improvement in filtering efficiency is obtained by use of a single rejection filter to remove the band of frequencies around the second harmonic. All other harmonics can be removed easily by a standard RC low-pass filter, and, as mentioned in connection with the half-wave discriminator, this part of the filter attenuates only the higher frequencies and has therefore a minimal effect on the signal band. Here again it must be kept in mind that the rejection filter must attenuate effectively all the frequencies in a band that is centered around the second harmonic of the carrier and has a width equal to twice the maximum frequency expected in the signal envelope.

The improved performance possible with the full-wave discriminator is obtained at the expense of considerable additional circuit complexity. This is easily appreciated by comparing Figs. 6.4 and 6.1. In order to get accurate full-wave output, the two halves of the circuit of Fig. 6.4 must be carefully balanced; i.e., the gain must be the same no matter which triode pair is conducting. If this is not the case, then in Fig. 6.5a, for instance, every other pulse will be larger than the two adjacent pulses (see Fig. 6.10a), and the output will then contain a component at the fundamental of the carrier frequency.

6.4. Other Triode Discriminator

Circuits. The circuits shown in Figs. 6.1 and 6.4 have been presented primarily as examples of practical circuits giving a half-wave or full-wave output. Other circuits operating on the same principles may, however, be devised and may be preferable, depending on the application. We shall consider briefly three of these circuits; others may be found in the literature.¹

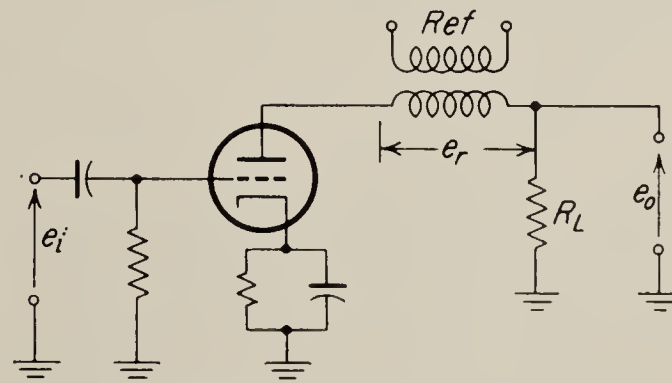


FIG. 6.7. Single-ended triode discriminator.

Probably the simplest triode circuit is the one shown in Fig. 6.7. This circuit represents essentially one half of the push-pull half-wave circuit

¹ A fairly complete list and references to other circuits are given in Greenwood, Holdam, and MacRae, "Electronic Instruments," Radiation Laboratory Series, vol. 21, McGraw-Hill Book Company, Inc., New York, 1948, pp. 383-386.

shown in Fig. 6.1. As in the other two circuits discussed, the important feature making this circuit a discriminator is the a-c plate supply, which disables the circuit every other half cycle. The output is a half-wave signal similar to that shown in Fig. 6.2, but since this is a single-ended circuit, the output voltage cannot reverse sign and is always negative during periods of tube conduction. The magnitude of the d-c component of the output varies continuously from a large negative value for a large, inphase signal input to smaller negative values as the input decreases to zero, reverses in phase, and increases again in the opposite direction. Hence, if it is necessary that the d-c output be zero for zero input, some sort of d-c potential-shifting device of the type described in Sec. 2.17 must be employed in the output. The output contains a large ripple component when the input signal is zero, and on the average the ripple is considerably larger in this circuit than in the push-pull circuit considered in Sec. 6.1. Thus a much larger amount of filtering is required to reduce the a-c component of the output to a permissible value. A further disadvantage of a single-ended circuit is that it has no drift compensation.

The analysis of this circuit proceeds in most essentials like that of the push-pull half-wave discriminator given in Sec. 6.2. The lowest ripple harmonic is the fundamental of the carrier, and therefore the upper limit on the envelope frequency is one-half the carrier frequency. The circuit operation cannot, however, be explained in terms of multiplication of the input signal by a square wave, since this would not account for the half-wave output obtained for zero input. A more reasonable assumption here is that the sum of the input and the sinusoidal reference is multiplied by the square wave. This assumption is justified by noting that the plate current of a triode may be expressed by

$$i_p = \frac{1}{r_p} (\mu e_g + e_p)$$

When the analysis is carried out on this basis, it will be found that both the quiescent d-c value and the magnitude of the ripple are determined primarily by the magnitude of the added reference voltage. This voltage should, therefore, be kept as small as possible, but on the other hand it must be large enough so that the sum of input and reference is always positive during the on period of the circuit. It is clear in summary that, although the circuit appears to be simpler than the one shown in Fig. 6.1, the extra filtering and d-c potential-shifting circuits that must be used with it largely nullify this apparent simplicity.

Both the circuits shown in Figs. 6.1 and 6.4 can be converted to cathode-follower circuits if it is not required that the signal be amplified and if the low output impedance and low drift afforded by cathode followers is desirable. In Fig. 6.8 is shown a half-wave cathode-follower circuit cor-

responding to the circuit of Fig. 6.1. Again its operation depends on the fact that the tubes are disabled by the reference signal every other half cycle. It can, therefore, be analyzed exactly as the circuit of Fig. 6.1 was, and its performance is identical except that the gain is slightly less than unity. The reason for grounding the reference transformer at the midtap of the winding is to provide a negative voltage for the cathode

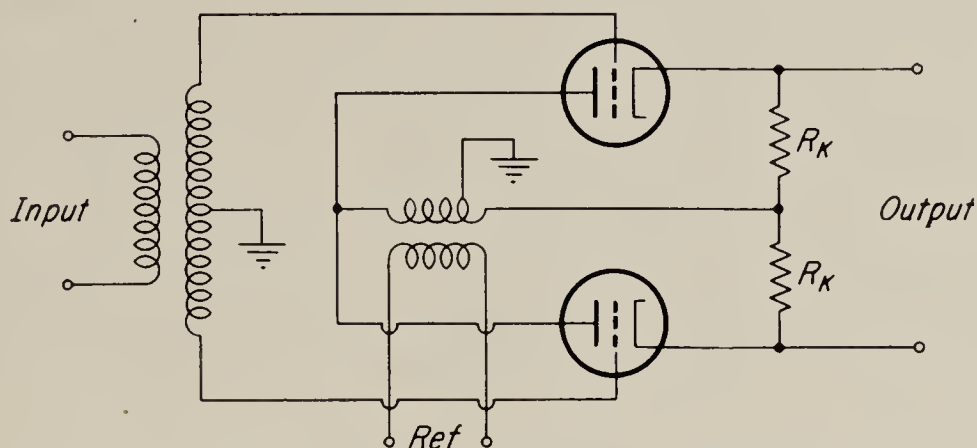


FIG. 6.8. Half-wave cathode-follower discriminator.

return during the operating half cycle. The placement of the tap need not be accurate. The full-wave counterpart of this circuit can be built up just as the full-wave circuit of Fig. 6.4 was built up from the half-wave circuit of Fig. 6.1. Construction of a detailed circuit is left as an exercise for the reader.

6.5. Diode Discriminators. All the triode discriminator circuits discussed thus far drift to a greater or lesser degree because they combine the functions of discriminator and d-c amplifier. For this reason diode discriminators are often preferred, particularly in systems having sufficient

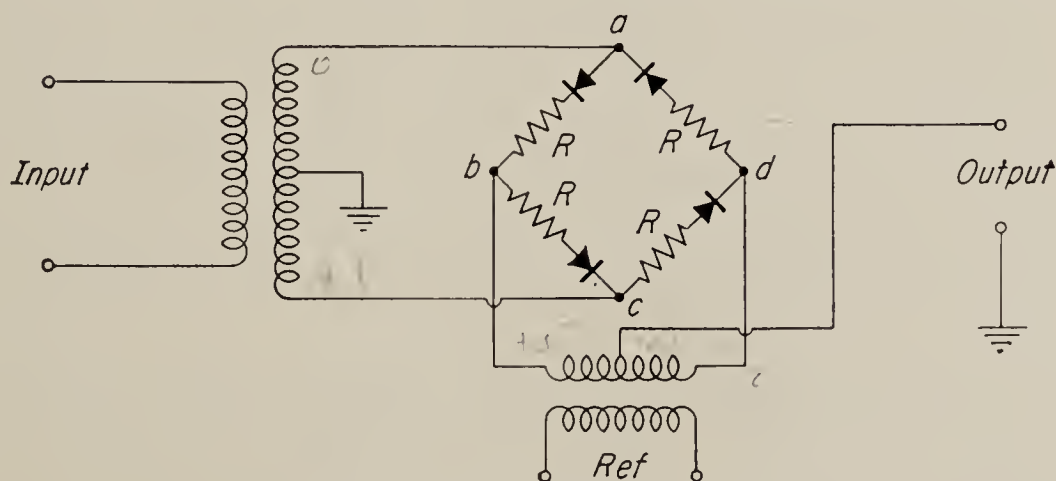


FIG. 6.9. Ring demodulator.

a-c amplification without additional d-c amplifiers. A practical advantage of diode discriminators is that the use of crystal diodes permits very simple and compact circuitry.

A popular diode discriminator circuit is the so-called *ring demodulator*, a schematic diagram of which is shown in Fig. 6.9. This circuit is a full-wave discriminator, and its operation, which is typical of all diode dis-

criminators, may be explained as follows: The reference signal is assumed to be very much larger than the input signal. Therefore, for the half cycle during which point b is positive and d negative, the lower two diodes conduct and the upper two are blocked. Hence there is an electric connection between the output terminal and point c , the lower end of the input transformer. If the secondary of the reference transformer is accurately center-tapped and if the impedances of the circuits bc and cd are exactly the same, there will be no voltage between point c and the output. The output voltage will therefore be the same as that of point c . During the next half cycle, point b is negative and d is positive. Thus the upper two diodes conduct, and the output is connected to point a . The circuit acts as a switch, connecting the output to either point a or point c in synchronism with the reference signal. The resistances R are required to limit the current flowing in the diodes and to equalize the impedances of the four arms.

From the description it should be clear that the operation of the circuit consists of multiplying the input signal by a square wave of unit amplitude (if the turns ratio of the input transformer is unity). Hence the analysis of the full-wave discriminator (Sec. 6.3) is directly applicable, and the output waveshapes observed under various operating conditions are ideally as shown in Fig. 6.5. It follows that the lowest ripple harmonic is the second harmonic of the carrier and that the maximum envelope frequency of the input is the carrier frequency.

In practice, unbalances between the bridge arms and inaccuracies in the location of the taps of the transformers result in output waveshapes that depart considerably from the ideal forms shown in Fig. 6.5. The effect of these unbalances is that the output contains a component at fundamental frequency, so that a filter with a reduced passband is required to smooth out the ripple. Since this to some extent defeats the purpose of the full-wave discriminator, some effort to balance the circuit properly is usually justified. The various adjustments required for this end are best understood by considering the various output waveforms in some detail.

Typical waveforms observed when the signal is in phase with the reference are shown in Fig. 6.10. The voltage of Fig. 6.10*a* may be obtained by adding a sine wave, at fundamental frequency and in phase with the reference, to a true full-wave-rectified signal like that shown in Fig. 6.5*a*. It can be shown by means of elementary circuit analysis that this type of output results from inaccurate midtap location on either of the two transformers or from inequalities in the resistors used in the bridge arms. If the center tap of the input transformer is incorrectly located, then no compensating adjustment of either the resistances or the reference center tap can be made to reduce this unbalance except at one signal amplitude.

On the other hand, if the signal transformer is not at fault, then it should be possible by the adjustment of any one of the four resistors to produce a true full-wave-rectified output. However, it may then be found that there is some d-c output for zero input. If this is undesirable, then another adjustment must be provided. Assuming that the transformer taps cannot be moved, one of the resistors in circuit *bad* and one in circuit *bcd* of Fig. 6.9 must be made adjustable to permit both zero d-c output for no input and a balanced full-wave output to be achieved.

The waveform shown in Fig. 6.10*b* results from the addition of a full-wave-rectified voltage and a sine wave at fundamental frequency but 90° out of phase with the reference. This effect cannot be produced by resistance unbalance but may be caused by unequal capacitances between the two halves of the reference winding and ground. This in turn may be caused by a reference transformer that does not have an electrostatic

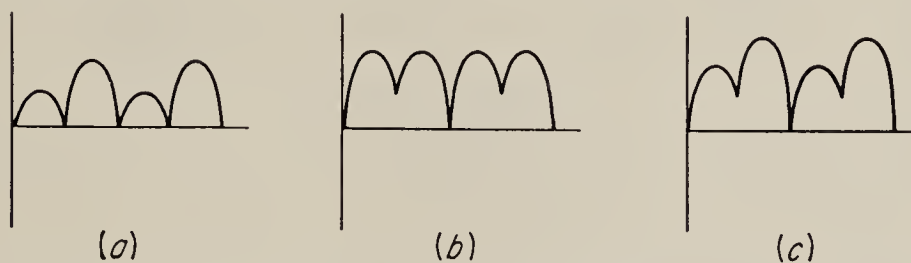


FIG. 6.10. Output waveforms resulting from circuit unbalance.

shield between its primary and secondary windings. The effect may be simulated by a small capacitor connected from either point *b* or point *d* to ground (Fig. 6.9). The remedy is to use a transformer with a Faraday shield or to add compensating capacitors between points *b* or *d* and ground. In Fig. 6.10*c* is shown the result when both the effects described occur simultaneously. Both resistive and capacitive adjustments must then be made to convert the output to the balanced full-wave form.

In addition to the balancing problem, another point of interest in this circuit is the maximum input signal for which the circuit will still function properly. To determine this, assume that the circuit is properly balanced and that the lower two diodes (Fig. 6.9) are conducting. The voltage across either one of the upper diodes can then be found by Kirchhoff's laws. Let e_o be the output voltage and e_a , e_b , etc., the voltage of points *a*, *b*, etc., all with respect to ground, and let e_R be the reference voltage across one half of the reference-transformer secondary [that is, $e_R = \frac{1}{2}(e_b - e_d)$]; then the voltage across the rectifier in arm *ab* is given by $e_a - (e_o + e_R)$. However, during the period when the lower set of rectifiers conducts, $e_o = e_c = -e_a$; hence the rectifier voltage is $2e_a - e_R$. Similarly the rectifier voltage in leg *ad* is given by $2e_a + e_R$. The reference voltage e_R is positive, since the lower rectifiers conduct. Hence, if e_a is positive, the voltage across the rectifier in arm *ab* becomes positive when e_a exceeds $e_R/2$, and the rectifier conducts. For negative e_a the

voltage across the rectifier in arm ad becomes negative when e_a exceeds $e_R/2$, and this rectifier conducts. Since neither of these rectifiers should conduct when the lower set conducts, this represents improper operation. Letting $e_a = e_s$, the signal voltage, we may therefore say that for proper operation

$$e_s < \frac{1}{2}e_R \quad (6.15)$$

Note that this computation has also given us the maximum back voltage on the rectifiers: for the maximum permissible signal this is equal to $2e_R$.

In applications permitting a half-wave-rectified output signal, simpler discriminator circuits like those shown in Fig. 6.11 may be used. The circuit shown in Fig. 6.11a is essentially half of the ring-modulator circuit

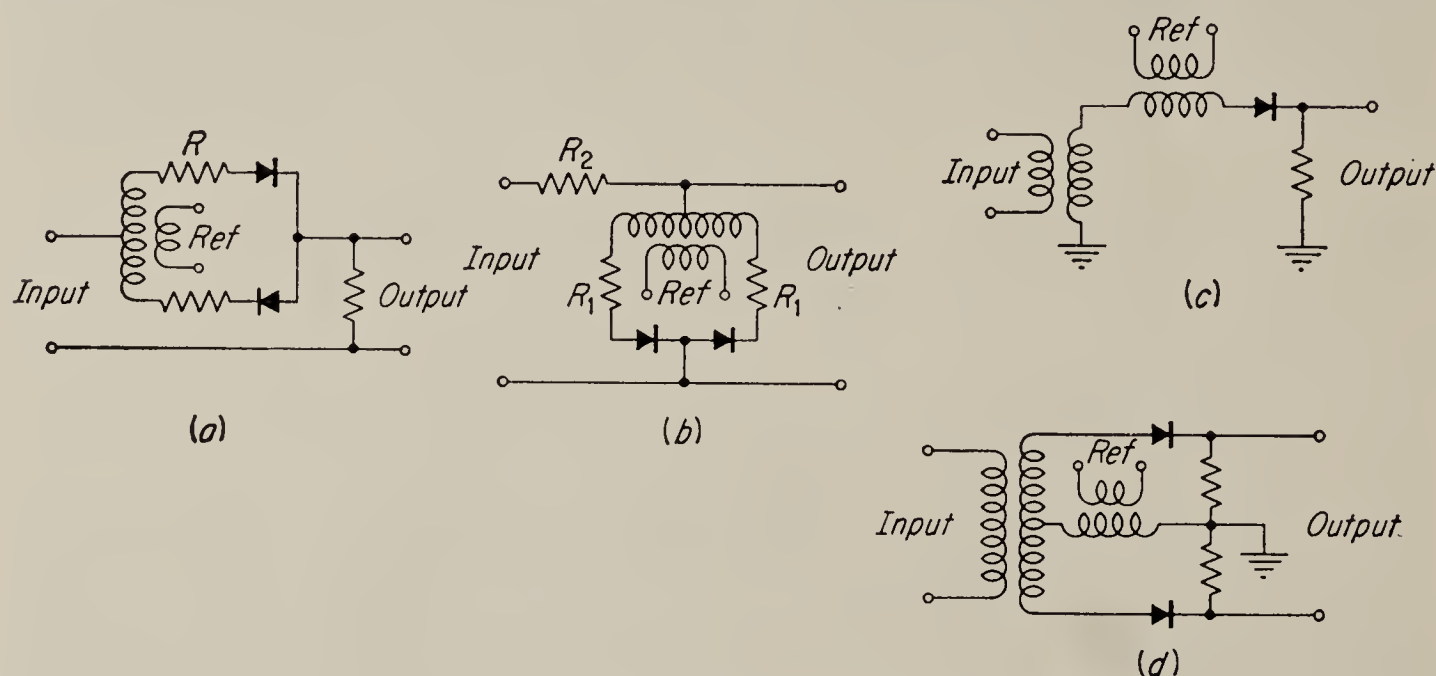


FIG. 6.11. Half-wave diode discriminators.

and functions the same way. During the half cycle when the diodes conduct, the output and input are connected together, and during the next half cycle, the output is zero. A slightly different version of this circuit is shown in Fig. 6.11b. Here $R_2 \gg R_1$, so that, during the periods when the diodes conduct, the signal is essentially shorted to ground. Both of these circuits when properly adjusted will have zero d-c and ripple output when the input is zero. This is not true in the circuit shown in Fig. 6.11c, whose output is always a half-wave-rectified wave. The amplitude of the output signal is increased by an input that is in phase with the reference and decreased by a signal that is out of phase with the reference. Thus, although this circuit is exceedingly simple, it requires a d-c potential-shifting circuit to make the average value of the output wave zero when the input is zero, and it usually requires more extensive filtering. Placing two circuits of the form of Fig. 6.11c back to back results in the push-pull circuit shown in Fig. 6.11d. The reader will recognize that the circuits of Fig. 6.11c and d are the diode counterparts of the triode circuits of Fig. 6.7 and Fig. 6.1, respectively.

A full-wave diode discriminator capable of handling relatively large currents, which may therefore be used to couple an a-c power amplifier to a load requiring a d-c signal, is shown in Fig. 6.12. The reference-transformer windings are connected in such a way that during one half cycle all the diodes in the upper diamond conduct and the ones in the lower diamond are blocked. During the next half cycle the situation is reversed. The circuit therefore acts as a switch that connects first the upper and then the lower end of the input transformer to the load in synchronism with the reference. The advantage of this circuit over some of the others that have been shown in this chapter is that the load current must flow only through the rectifiers and therefore encounters only their forward resistance. Hence this circuit can be built with signal efficiencies (ratio of output power to power supplied by signal) of approximately 80 per cent. The

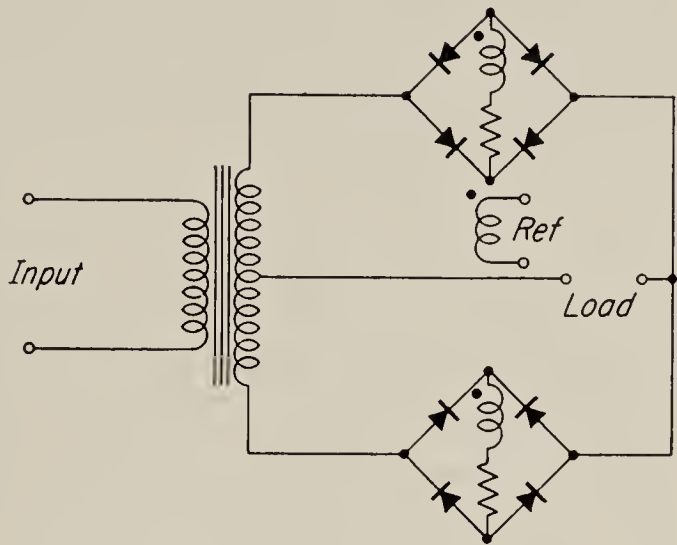


FIG. 6.12. Diode power discriminator.

current rating is a function primarily of the current rating rectifiers and of the reference transformer, but it can easily reach the order of amperes.

6.6. Electromechanical Demodulators. Most of the diode circuits described in the previous section were shown to operate essentially as synchronous switches. Hence it is clear that a mechanical synchronous switch could act as a demodulator in the same way as the electronic circuits described. Switches of this sort, referred to as *electromechanical vibrators*, or *choppers*, are available commercially and are able to handle frequencies of up to about 1,000 cps. One application of such a device has

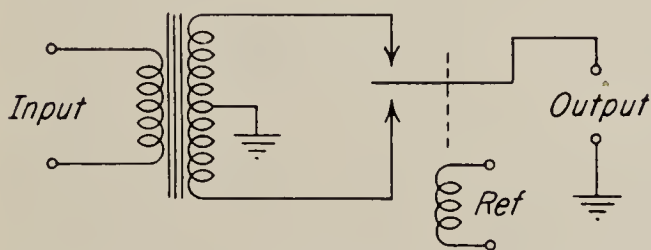


FIG. 6.13. Demodulator using mechanical chopper.

already been described in Sec. 2.23 and the reader is referred to that section for a short description. Figure 2.34 shows the external appearance and internal connections of a typical commercially available unit. A typical demodulator circuit using choppers is shown in Fig. 6.13; more elaborate circuits can be devised if the application warrants it.¹ The output of the circuit shown is a full-wave signal; hence the output waveshapes are of the sort shown in Fig. 6.5. In order to minimize the amount of driving power required, the central reed is usually constructed to have

¹ James, Nichols, and Phillips, "Theory of Servomechanisms," Radiation Laboratory Series, vol. 25, McGraw-Hill Book Company, Inc., New York, 1947, p. 109.

needed in order that the charging current may flow in either direction. If the amplitude of the reference signal is very large compared to the cutoff bias voltage of the triodes, the period of tube conduction is very short, and it may be assumed that the charge placed on the capacitor C_2 represents the value of the input signal at the instant that the reference passes through its peak; i.e., the circuit *samples* the voltage existing at the input at this instant. The load impedance connected to the circuit is assumed to be infinite. Hence the capacitor holds, or "clamps," the signal at this level until the application of the next positive reference-voltage peak. A typical form of the output for a signal in which the carrier is in phase with the reference is shown in Fig. 6.15. Note that a constant-amplitude

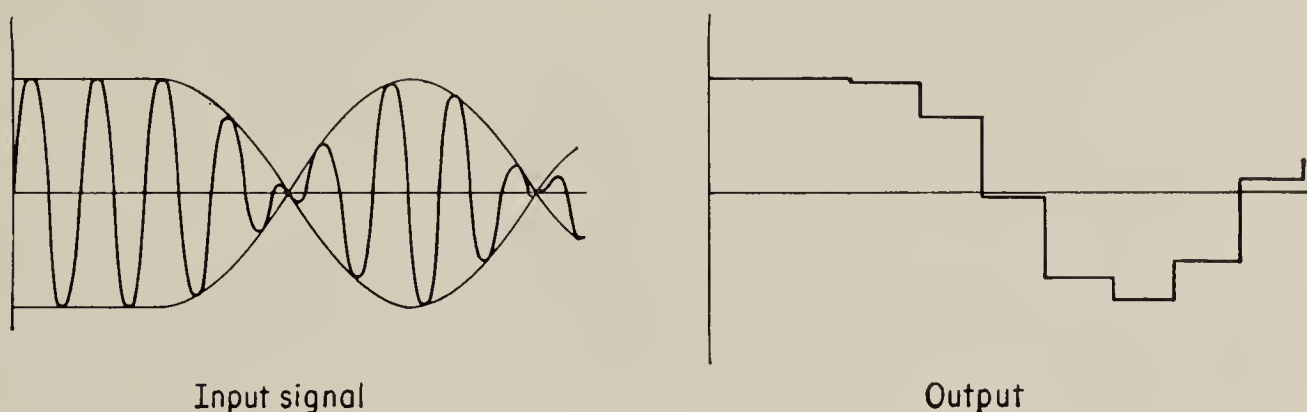


FIG. 6.15. Output waveshapes for typical input.

input ideally results in a d-c output, while for a varying input the output has a steplike appearance. The ripple content of the output is much less than for more conventional discriminator circuits: To appreciate this fully, the reader should keep in mind that the circuit described here is a half-wave circuit; i.e., the signal is sampled only once per cycle.

In the circuit design, the RC time constant of the grid-leak circuits should be of an order of magnitude larger than the period of the reference signal in order that the capacitor shall not discharge significantly during one period. The grid-leak resistor must be large enough to prevent excessive grid current from being drawn at any time. The output capacitor C_2 should have such a value that the RC time constant of this capacitor in combination with the resistance of the conducting tubes will be very short compared to one period of the reference signal so that the capacitor will charge up to the full signal level during the short time that the tubes conduct. On the other hand the time constant with the tube not conducting should be very much greater than a period of the reference to ensure that the signal is properly held during the time between sampling instants. Hence the load impedance should be several orders of magnitude greater than the plate resistance of the triodes, and a cathode follower is preferably used here. It is clear that the improvement in performance is obtained at the expense of a somewhat more complicated circuit. This is particularly true if, as is commonly done, the reference is

passed through *peaking circuits* and applied to the discriminator in the form of sharp pulses so as to shorten the sampling time.

If the carrier component of the input signal is 90° out of phase with the reference, the input signal at the sampling instants is always zero, and hence the output is zero. In general, if the phase difference between input and reference is greater than 0° but less than 90° , the output amplitude is reduced in proportion to $\cos \beta$, where β is the phase difference. In this respect, therefore, the circuit acts like a standard phase-sensitive discriminator.

Since the circuit is a true sampler, the limiting frequency of the envelope of the input is equal to one-half the number of samples taken per

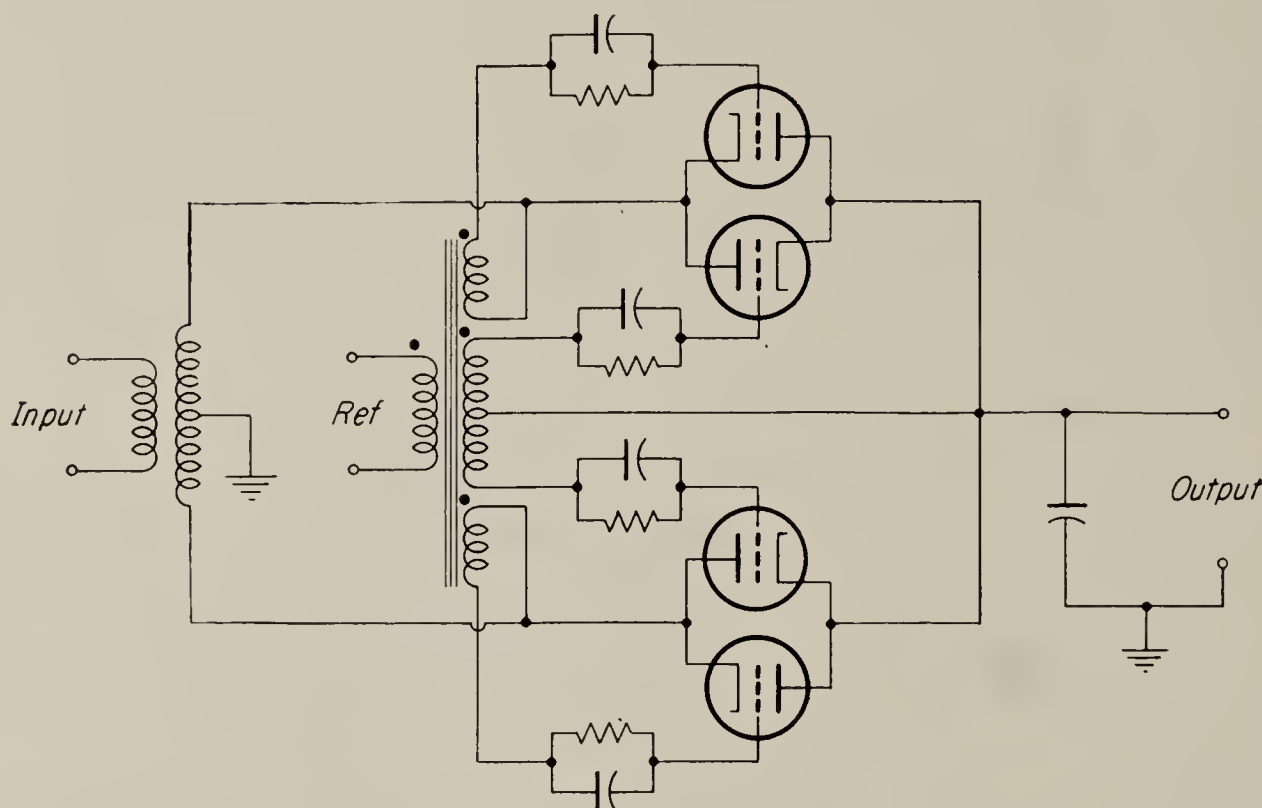


FIG. 6.16. Full-wave clamping discriminator.

second.¹ Hence an increase in the sampling rate makes the circuit useful for higher-frequency signals, and it also reduces the amount of ripple at a given input frequency. One way to double the sampling rate is to use the full-wave equivalent of the clamping discriminator, shown in Fig. 6.16. The connection of the reference transformer is such that when the upper pair of triodes conduct, the lower pair is cut off, and vice versa. The operation of the triodes is the same as already described. If we suppose the upper end of the input transformer to be positive when the upper pair of triodes conduct, a positive charge is placed on the output capacitor; during the next half cycle the lower half of the transformer is positive and the lower pair of tubes conduct, placing another positive charge on the capacitor. Thus two samples per reference cycle are obtained.

By modifying the reference signal, the circuit of Fig. 6.16 can be

¹ Truxal, "Automatic Feedback Control System Synthesis," McGraw-Hill Book Company, Inc., New York, 1955, pp. 500ff.

arranged to yield four samples per reference cycle. This is done by applying a reference pulse to the tubes at the instants that the carrier, assumed to be in phase with the reference, passes through the 45° and 135° points (see Fig. 6.17). Obviously a fairly elaborate shaping circuit would be required to obtain the required reference signal, and the additional complication would probably be worth while only if several discriminators could all be operated from one reference source in a particular system.

A very simple method for determining the frequency response of a clamping discriminator is given by Truxal,¹ who shows that the transfer function may be expressed as

$$H(s) = \frac{1 - e^{-Ts}}{Ts} \quad (6.16)$$

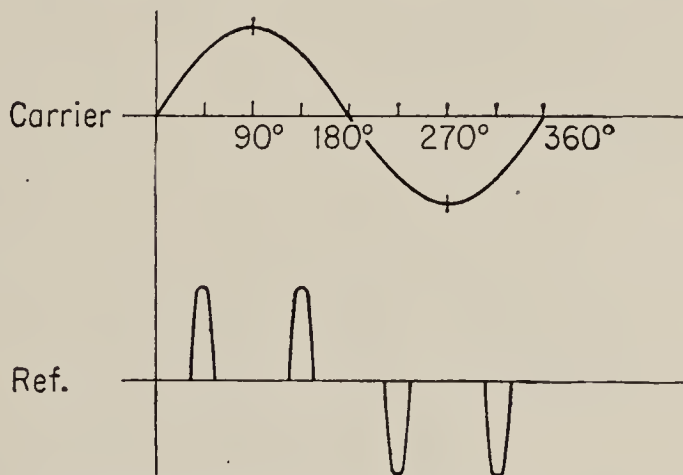


FIG. 6.17. Relation of reference to carrier to provide four samples per reference period.

where T is the period between samples. The gain and phase shift as a function of frequency of the input envelope, ω_s , can be obtained from this expression by setting $s = j\omega_s$ and simplifying the resulting complex function. The result may be put in the form

$$H(j\omega_s) = \left| \frac{\sin(T\omega_s/2)}{T\omega_s/2} \right| e^{-jT\omega_s/2} \quad (6.17)$$

or, if we let the sampling frequency be $\omega = 2\pi/T$, this may be written in the equivalent form

$$H(j\omega_s) = \left| \frac{\sin \pi(\omega_s/\omega)}{\pi\omega_s/\omega} \right| e^{-j\pi\omega_s/\omega} \quad (6.18)$$

The absolute values of the gain and the phase shift are sketched in Fig. 6.18. In practice, only the portion below $\omega/2$ is of interest, since $\omega/2$ is the limit of unambiguous transmission. The gain at this limit is $2/\pi$, and the phase lag $\pi/2$ radians. Note that the sampling frequency ω is equal to the reference frequency for a half-wave clamping discriminator and to twice the reference frequency for a full-wave discriminator. Thus the results for the upper limit of transmission obtained here are identical with those obtained with the simpler discriminator circuits discussed at the beginning of this chapter.

In order to find the magnitude and frequency of the ripple harmonics in the output, a Fourier analysis must be made on a typical output, such as that obtained, for instance, by an input with a sinusoidal modulation.

¹ *Ibid.*, p. 507. Note that Eq. (6.16) is Truxal's expression multiplied by $1/T$. This is done to make the gain unity for zero frequency.

Such a Fourier analysis can be made most simply by first finding the Laplace transform of the output and determining the harmonics from this by partial-fraction expansion.

The Laplace transform of the output signal obtained for a sinusoidal input modulation of frequency ω_s is derived by Gardner and Barnes¹ in connection with their discussion of the solution of difference equations.

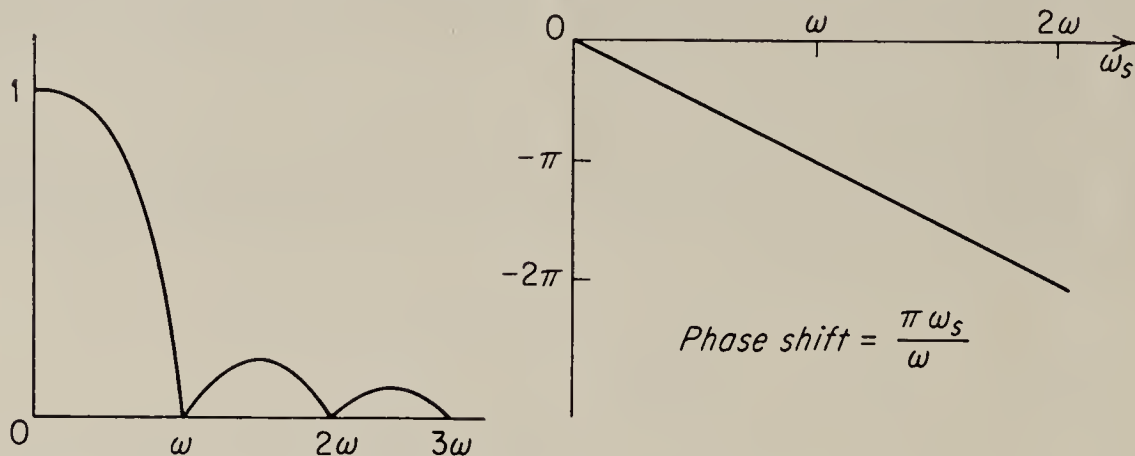


FIG. 6.18. Gain and phase shift of clamping discriminator.

It may also be obtained by application of the theory of sampled data systems.² In either case it is given by

$$F(s) = \frac{(\sin \omega_s T)(e^{sT} - 1)}{s(e^{2sT} - 2e^{sT} \cos \omega_s T + 1)} \quad (6.19)$$

The poles of this function are found by solving for the zeros of the denominator; they occur for $s = 0$ and $s = j(2\pi n/T \pm \omega_s)$, $n = 0, \pm 1, \pm 2, \dots$. Hence $F(s)$ may be written in the equivalent form

$$F(s) = \frac{k_0}{s} + \sum_{n=-\infty}^{\infty} \left\{ \frac{k_n}{s + j[(2\pi n/T) + \omega_s]} + \frac{k'_n}{s + j[(2\pi n/T) - \omega_s]} \right\} \quad (6.20)$$

Since $2\pi/T = \omega$, the sampling frequency, it may be seen from Eq. (6.20) that the harmonic frequencies are $n\omega \pm \omega_s$, a result completely analogous to that obtained previously for the simple half-wave discriminator [Eq. (6.7)]. The k_n may be evaluated by standard techniques which, although fairly cumbersome, offer no difficulties in principle. The result of Eq. (6.20) may then be put into the form

$$F(s) = \sum_{n=-\infty}^{\infty} \left\{ \left(\frac{\sin \omega_s T}{2\pi n + \omega_s T} \right) \left(\frac{(2\pi n/T) + \omega_s}{s^2 + [(2\pi n/T) + \omega_s]^2} \right) + \left(\frac{\cos \omega_s T - 1}{2\pi n + \omega_s T} \right) \left(\frac{s}{s^2 + [(2\pi n/T) + \omega_s]^2} \right) \right\} \quad (6.21)$$

¹ Gardner and Barnes, "Transients in Linear Systems," John Wiley & Sons, Inc., New York, 1947, p. 299.

² Truxal, *op. cit.*, pp. 507-517.

The second fractions of the two products are seen to be the transforms, respectively, of $\sin [(2\pi n/T) + \omega_s]t$ and $\cos [(2\pi n/T) + \omega_s]t$. Hence the amplitude of the ripple component at $(2\pi n/T) + \omega_s$ is obtained by taking the square root of the sum of the squares of the coefficients of these two factors, with the final result for the amplitude of the n th harmonic:

$$A_n = \frac{\sin (\omega_s T/2)}{\pi n + (\omega_s T/2)} \quad (6.22)$$

for the harmonic frequency $(n\omega + \omega_s)$; and

$$A'_n = \frac{\sin (\omega_s T/2)}{\pi n - (\omega_s T/2)} \quad (6.23)$$

for the harmonic frequency $(n\omega - \omega_s)$.

Note that for $n = 0$ this result reduces to the amplitude of the desired output, obtained already by a different argument [Eq. (6.18)]. Note also that, as $\omega_s \rightarrow 0$, the harmonics go to zero; this agrees with our previous finding that a constant-amplitude input yields no ripple output.

6.8. Modulators. Modulators are used in servo systems when it is desirable to convert the low-frequency signals to a higher frequency. A common example is a servo system in which an a-c motor is used in the power stage, yet where the error signal is a d-c (low-frequency) voltage. In such a system the error is usually modulated at a low signal level, and the resulting high-frequency signal is amplified in a-c amplifiers and applied to the motor. In this way the drift problems encountered with d-c amplifiers are avoided. Sometimes both demodulators and modulators are used in a system, even though both the error signal and the voltage required by the power unit are alternating current. This is done because the loop-transfer shaping networks or equalizers that are required to obtain optimum system performance can be given much more satisfactory characteristics when they are designed to operate in the low-frequency-signal spectrum than when they must operate on the modulated signal directly. Hence, where such networks are to be used, the signal coming from the transducer is first passed through a demodulator, then operated upon by the network, then remodulated, amplified, and applied to the motor. A further example of the use of modulators and demodulators is the drift-free d-c amplifier described in Chap. 2 (Sec. 2.23). Since a-c servomotors and most of the demodulators used in servomechanisms require a suppressed-carrier modulation and also since the suppressed-carrier modulation is more symmetrical than standard amplitude modulation (i.e., the amplitude of the a-c output is independent of the sign of the input), almost all modulators designed for use in servo systems yield suppressed-carrier type of modulation.

All the demodulator circuits discussed in the preceding sections, except

for the clamping discriminator, will function as modulators if the input is a low-frequency signal rather than a modulated carrier. This is so because, as has been explained previously, all of these circuits operate to multiply the input by a square wave. Hence, when a low-frequency input signal is applied, for instance, to the ring-demodulator circuit (Fig. 6.9), the modulated output has the appearance shown in Fig. 6.19. The carrier is a square wave; this is advantageous if the signal is to be demodulated by another square-wave multiplication, since the ripple output is much less than for a sine-wave carrier. On the other hand, if the signal

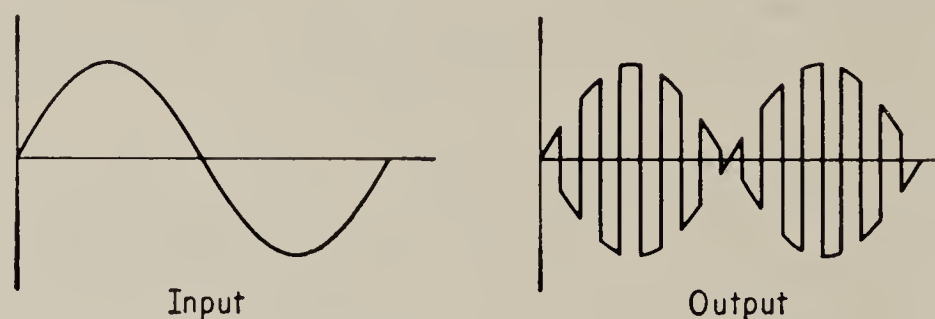


FIG. 6.19. Output of full-wave modulator.

is to be applied to the control phase of an a-c motor, the square-wave carrier is not desirable; the higher harmonics may, however, be removed by low-pass filters.

An important modification is necessary on circuits such as that of Fig. 6.9, if, as is usually the case in servo applications, the modulator must be able to handle input frequencies down to zero frequency. An input transformer cannot be used under these conditions to provide the push-pull input required in the operation of the circuit, and a direct-coupled phase inverter (see Chap. 2, Secs. 2.15 or 2.16) would be required instead.

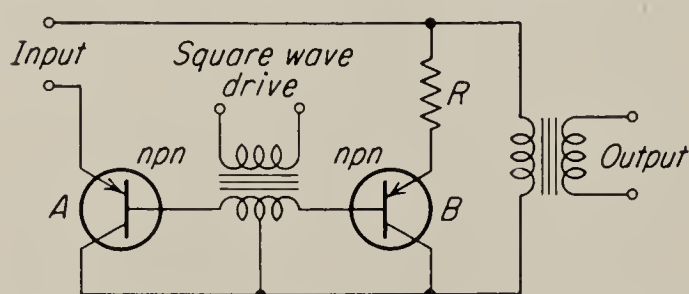


FIG. 6.20. A simple transistor chopper.
(Bright, Kruper)

This complication is, however, avoided with the circuit of Fig. 6.9 by operating it backwards, i.e., by applying the low-frequency input at the terminal used for the output when the circuit is used as a demodulator and taking the a-c output from the transformer terminals.

The reader may show that the operation of the circuit is not affected by this procedure.

Very successful low-level square-wave modulators can be constructed with transistors.¹ Typically, *n-p-n* fused-junction transistors have been used, but *p-n-p* types could also be employed. The major advantages of transistor choppers are (1) no appreciable phase shift even with chopping

¹ A. P. Kruper, Switching Transistors Used as a Substitute for Mechanical Low Level Choppers, *Trans. AIEE*, vol. 74, part 1, pp. 141-144, 1955. R. L. Bright, Junction Transistors Used as Switches, *Trans. AIEE*, vol. 74, pp. 111-121, 1955.

rates as high as 10 kc and (2) excellent linearity for signals down to 0.1 millivolts in amplitude. There is little temperature drift so long as the input impedance is kept low. Figure 6.20 shows a simple transistor chopper.¹ The square-wave drive applied to the bases of the transistors alternately switches one transistor on and the other off. For the half cycle in which *A* conducts, *B* is cut off; thus the input is connected to the output through the output transformer. For the other half of the cycle, *A* is cut off and the path from the input is open-circuited. At the same time *B* conducts and shorts the output through resistor *R*. While a common emitter circuit could be used here, the common collector was chosen because the static collector voltage for zero collector current is only about 1 mv. This voltage appears across the output transformer in the same polarity for both transistors, and if the transistors are matched, this is a constant d-c value and thus does not appear at the output terminals.

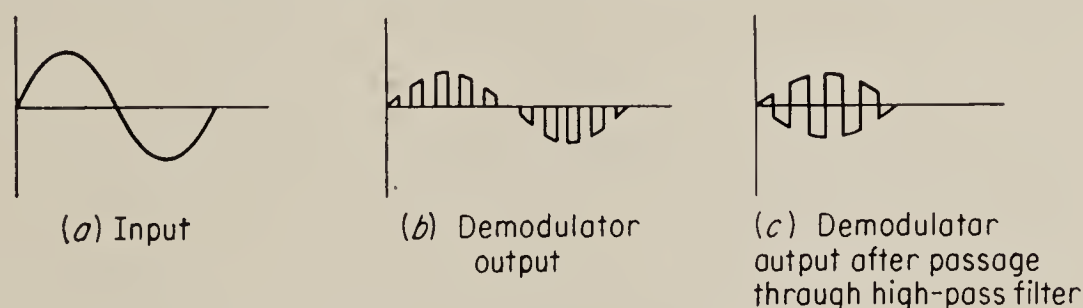


FIG. 6.21. Waveshapes obtained from half-wave modulator.

When a half-wave demodulator is used as a modulator, the low-frequency signal is multiplied by a square wave having a d-c value equal to one-half the peak-to-peak square-wave amplitude (see Sec. 6.2). Hence the output is as shown in Fig. 6.21b, and it contains a low frequency as a harmonic component. If this signal is amplified, however, in an audio amplifier having the usual high-pass coupling networks, then the low-frequency component is removed from the output, and the signal shown in Fig. 6.21c results. It is seen to be identical with that produced by the full-wave modulator (Fig. 6.19). Thus, except that the gain of the half-wave modulator is only one-half as large as for an equivalent full-wave modulator, there is no particular advantage in the use of a full-wave rather than a half-wave modulator.

6.9. Drift in Modulators. All electronic modulators drift a certain amount, with triode circuits being generally worse than diode circuits. The reasons for drift in triode circuits are simply shown. Consider the circuit shown in Fig. 6.22, which is essentially the same circuit as the demodulator of Fig. 6.1 except that it also acts as a phase inverter for the single-ended input signal (see Sec. 2.16). The operation of this circuit has been explained as multiplication by a square wave of amplitude $G/2$,

¹ R. L. Bright and A. P. Kruper, Transistor Choppers for Stable DC Amplifiers, *Electronics*, April, 1955, pp. 135-137.

where G is the gain during the conducting period. Clearly if, because of variation of tube parameters, G changes, the output for a given input changes. In particular, if the input is zero and the tube characteristics are not exactly identical, some square-wave output will result. Thus if the characteristics of one tube change relative to those of the other, the amplitude of this square-wave output can change; this constitutes drift.

For the diode circuit the cause of drift is somewhat different and depends on the fact that a diode will conduct some current when reverse voltage is applied. Hence, if we consider the half-wave circuit of Fig. 6.11a, we find that during the period of nominal nonconduction some signal may get to the output, which instead of oscillating between the signal amplitude and zero will instead oscillate between full-input amplitude and reduced-input amplitude. It is clear that, if temperature changes or other factors cause the back current of the diodes to change, then the

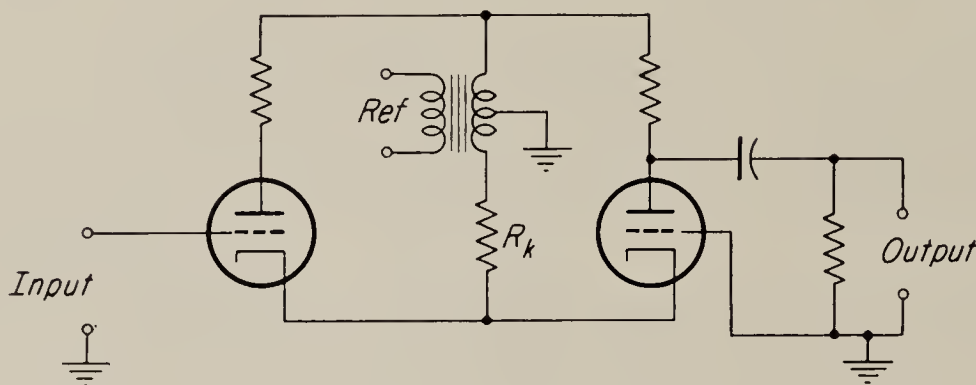


FIG. 6.22. Triode modulator.

magnitude of the output square wave changes for a given input, and we again have drift.

Since a major reason for the use of a modulator in a servo is to prevent the drift that is caused by d-c amplifiers, a modulator that drifts is of relatively little usefulness. For this reason the most commonly used modulator, particularly for systems using a 60- or 400-cps carrier, is the electromechanical chopper, already discussed previously (Secs. 2.23 and 6.6). It may be connected as shown in Fig. 6.13, but with the input and output terminals reversed, or it may be connected as a half-wave modulator, as shown in Fig. 2.35. In either case its good square-wave output and lack of drift are the primary advantages.

PROBLEMS

6.1. The output from a full-wave discriminator is passed through a notch-type rejection filter which completely removes the second harmonic of the reference frequency but has negligible effect on any other harmonics. Assume that the input signal is in phase with the reference. Plot the form of the output signal delivered by the discriminator and the notch network.

6.2. Repeat Prob. 6.1 for the case of (a) an input signal 90° out of phase with the reference, and (b) an input signal 45° out of phase with the reference.

6.3. A two-section low-pass filter used to smooth out the ripple generated by a half-wave discriminator operating on a 60-cps carrier has the transfer function

$$H = \left(\frac{1}{T_s + 1} \right)^2$$

It is desired to reduce the rms value of ripple output to less than 10 per cent of the useful d-c output. Assume the input signal is in phase with the reference and find the time constant T required.

6.4. Repeat Prob. 6.3 for a full-wave discriminator. Find the ratio of time constants required in the two discriminator types.

6.5. Design a ring demodulator like that in Fig. 6.9. Assume both transformers have unity ratio. The maximum expected value of the input signal is 20 volts rms. The diodes are 1N34, having a maximum inverse voltage rating of 60 volts and a maximum current rating of 50 ma. The circuit operates into a load resistance of 10,000 ohms, and it is desirable to keep the output impedance of the discriminator to a minimum. Find the values of the resistors R required.

6.6. Find the current rating of the secondaries of the reference transformer and of the rectifiers required in the power discriminator of Fig. 6.12. Express the result as a ratio of maximum load current.

6.7. In the clamping discriminator shown in Fig. 6.14 the triodes have a plate resistance of 1,000 ohms when they conduct. The pulses applied to the grids cause the tubes to conduct for one-twentieth of a cycle. During this time the capacitor should charge to 95 per cent of the correct input value. The capacitor should hold the correct value with a loss of less than 5 per cent during the period that the tubes do not conduct. The pulse-repetition frequency is 400 cps. Find the value of C_2 required, and also the minimum value of load resistance.

CHAPTER 7

A-C MOTORS

7.1. Introduction. The motor used most commonly in low-power applications is the a-c two-phase induction motor. When it is used as a servomotor, one of the phase windings is connected to a fixed a-c voltage, the *reference voltage*, and the other winding is supplied by a variable *control voltage*, 90° out of phase with the fixed voltage. When the control voltage leads the fixed voltage, rotation is in one direction; when it lags the fixed voltage, the direction of rotation is reversed. The speed and torque depend on the magnitude of the control voltage, and although the relation is not linear, it is sufficiently regular so that approximately proportional control may be achieved.

One of the reasons for the popularity of this type of motor is that it requires only the simplest type of control amplifier. A typical arrangement is shown in Fig. 7.1, where a synchro pair is used to indicate the

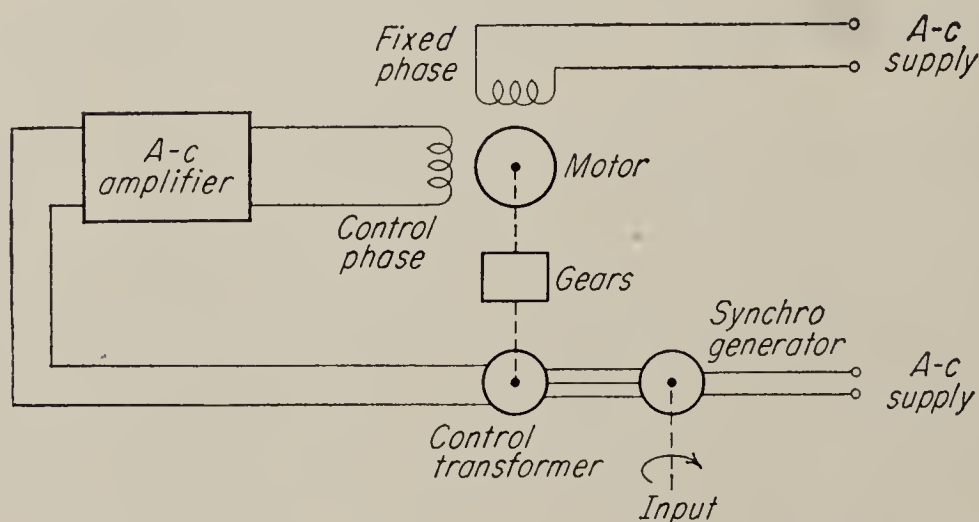


FIG. 7.1. All a-c servo using a two-phase motor.

error between input and output. The error voltage from the control transformer is amplified in a standard a-c amplifier and applied to the control winding of the motor. The 90° phase shift between the control and reference voltages may be obtained either by connecting a capacitor in series with the reference winding or by incorporating a simple phase-shift network in the amplifier. Neither method offers any particular difficulty, but the second has the advantage that the phase shift is not affected by the motor operating conditions. Another small departure

from conventional amplifier design results from the highly inductive input impedance of a-c motors. To improve the effective load power factor, a *resonating capacitor* is usually connected across the amplifier output terminals.

It should be noted that it is not necessary to pass the signal from the control transformer through a phase discriminator; the motor acts as its own discriminator. Hence many simple control systems require no d-c amplification at any point in the loop, and therefore the problem of drift associated with d-c amplifiers is eliminated.

Another desirable feature is that the motor, in common with all induction motors, requires no electric connections, such as brushes or slip rings, between its stator and rotor, and friction is thus reduced to an absolute minimum. This results in very smooth operation and is one of the reasons why the induction motor is used even in systems where the error signal is d-c and must be passed through a chopper or other modulator to actuate the motor.

A-c servomotors are built for power outputs ranging from $\frac{1}{2}$ to 1,000 watts; however they are most commonly found in applications requiring less than 10 watts. The efficiency is usually quite low, 5 to 20 per cent being typical. This, in addition to the fact that the motors often must operate near zero speed for extended periods of time, necessitates the use of an external blower on units of more than 10-watt output rating to provide proper cooling. Such a blower is usually driven by a separate single-phase induction motor.

7.2. Construction Features. The stator frames of a-c servomotors are all constructed in essentially the same way although motors of different manufacturers have external differences, and various mounting styles are available. Several types of commercially available servomotors are shown in Fig. 7.2. The stators have a standard distributed winding to improve the space distribution of flux density around the air gap. It is desirable to have a flux distribution as close to sinusoidal as possible, because space harmonics in the flux tend to produce notches and other irregularities in the speed-torque curve of the motor. Most 60-cps motors are wound for two or four poles, but 400-cps motors may be wound for as many as 10 or more poles, to reduce the operating speed. The two phases of the winding may be identical; however, in some motors the control winding is designed for a much higher voltage than the reference winding, and in some cases the control winding is center-tapped. These special features make it possible to connect the motor directly to the output stage of the amplifier without the need of an output transformer. It is also possible, particularly in the smaller motors, to design the stator winding in such a way that only a small fraction of the power input to the motor is supplied by the control amplifier,

the major share of the power coming from the reference supply. All of these special design features serve to reduce the size and weight of the control amplifier.

There are three basically different rotor designs: the *squirrel-cage* rotor, the *solid-iron* rotor, and the *drag-cup* rotor. Squirrel-cage rotors are quite similar to those used in standard induction motors except that, in order to reduce the moment of inertia, the diameter is usually much smaller than the length (see Fig. 7.3). Squirrel-cage windings are subject

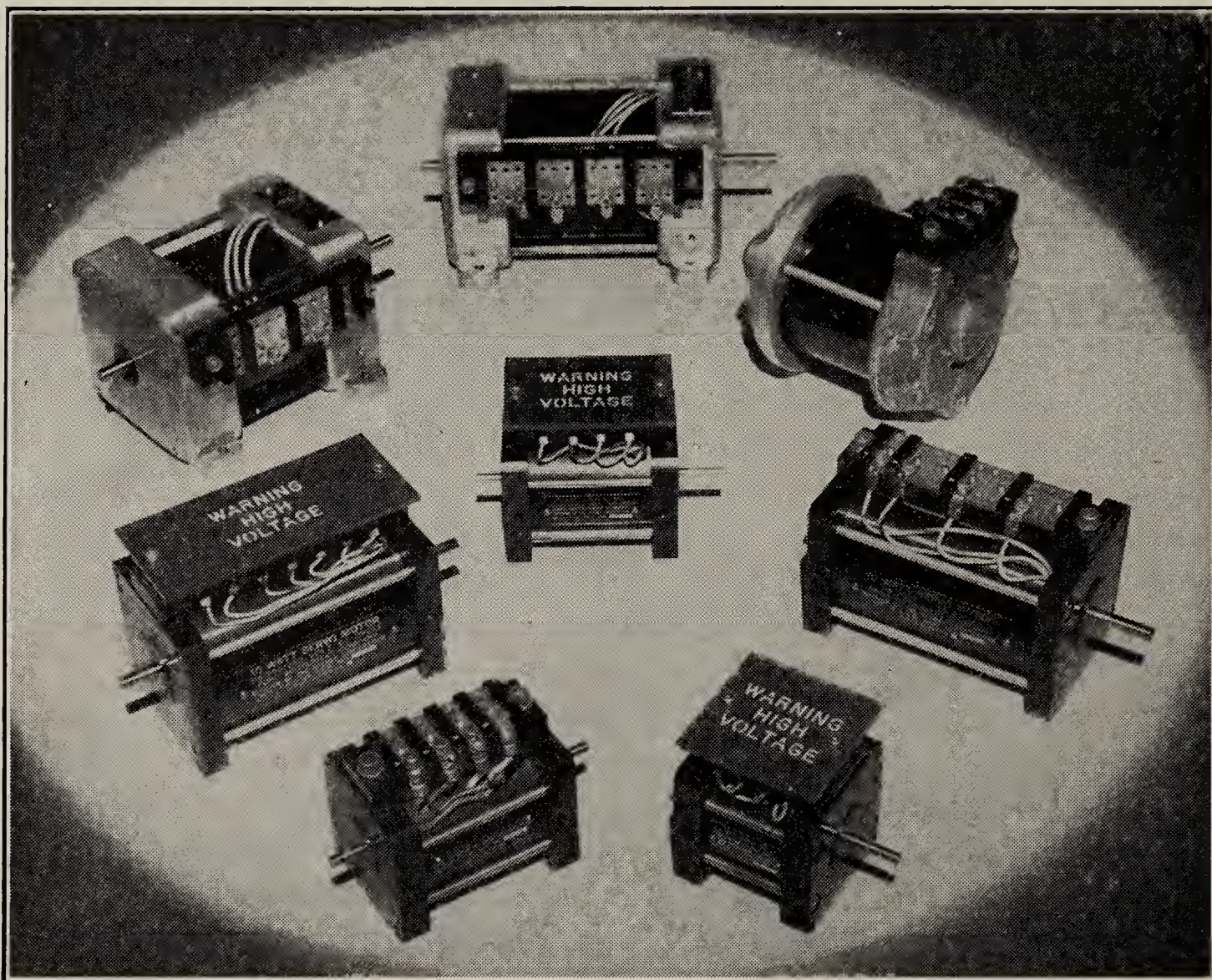


FIG. 7.2. Various types of servomotors. (Courtesy Ford Instrument Co.)

to *cogging*, or *slot lock*, a magnetic attraction between stator and rotor which keeps the rotor from turning until the control voltage reaches a certain minimum breakaway value. The winding is therefore usually skewed (see Fig. 7.3), to minimize this effect.

The solid-iron type of rotor is very similar in external appearance to the squirrel-cage type, being a slender cylinder. The iron used in its construction, in addition to having good magnetic qualities, must have a relatively high conductivity so that sufficient current can be made to flow in it. There is no slot-lock problem with this type of rotor; however, the torque developed is somewhat less than for an equivalent squirrel-cage winding.

Where the lowest possible moment of inertia is desired, a motor with a drag-cup type of rotor is used (see Fig. 7.4). The stator of a motor designed for this rotor has, in addition to the usual stator punchings and winding, a central cylinder also made of steel punchings to complete the magnetic circuit. The drag cup fits into the air space between the windings and the stationary cylinder, and the clearances are kept as small as possible to reduce the length of the air gap. Despite this, the air gap in a drag-cup motor is always very much longer than in motors using either of the other two rotors, and for this reason the torque developed by a

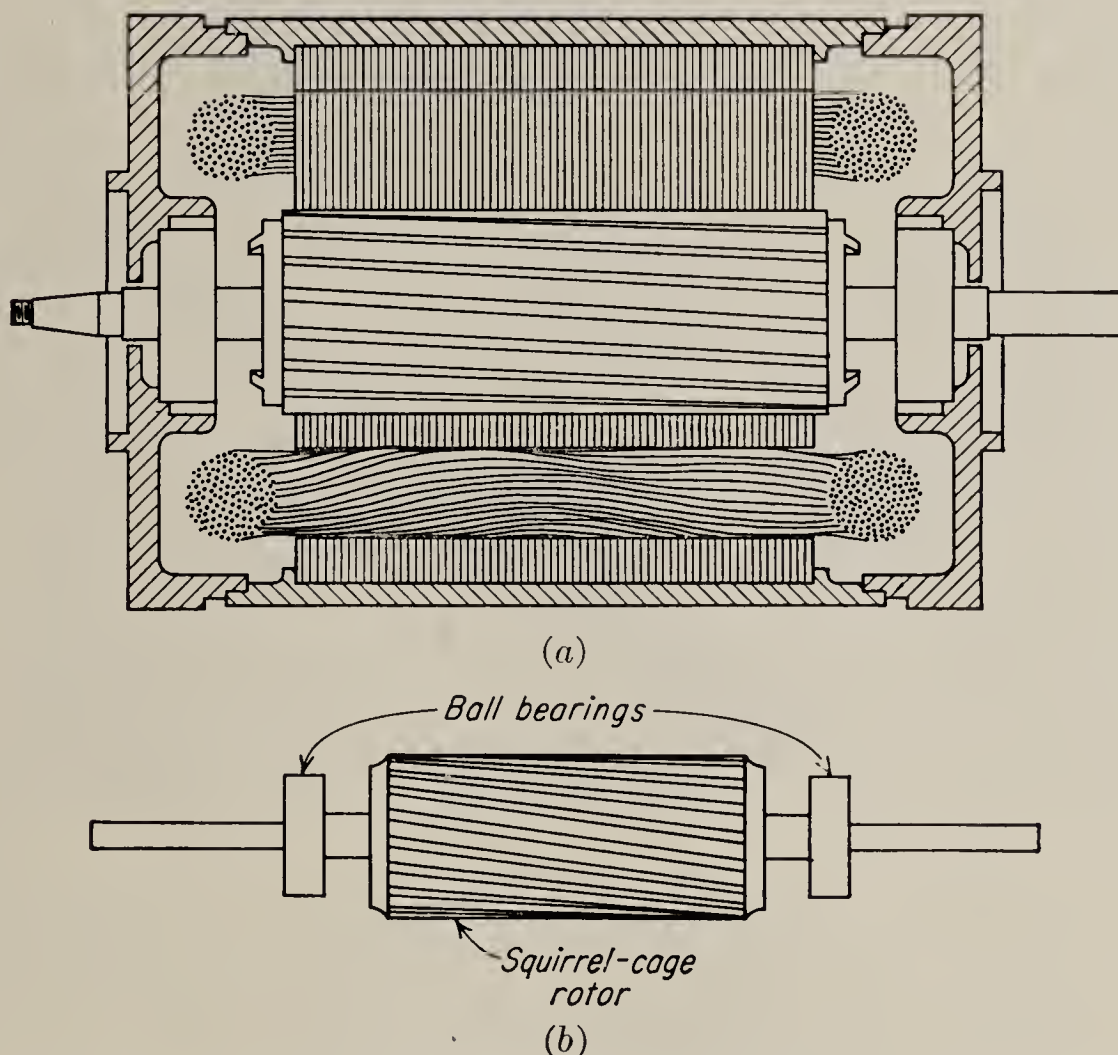


FIG. 7.3. Cutaway view of a-c servomotor with squirrel-cage rotor. Note long slim design and skewed rotor slots.

drag-cup motor is relatively small. Since, however, the inertia of the drag cup is extremely low, both the ratios of torque to inertia and of torque squared to inertia in the drag-cup motor compare favorably with those of the other rotors.

7.3. Theory of Operation. The theory of operation of the two-phase servomotor depends on the theory of the polyphase induction motor. We review this theory here briefly for the benefit of any reader not familiar with it. For a more detailed discussion of the theory of polyphase induction motors, the reader is referred to any one of the standard texts on a-c machinery.¹

¹ For instance, Puchstein, Lloyd, and Conrad, "AC Machines," 3d ed., John Wiley & Sons, Inc., New York, 1954, pp. 252-324.

A schematic diagram of a two-phase motor is shown in Fig. 7.5. The rotor is assumed to have a squirrel-cage winding, and there are two stator windings displaced in space by 90 electrical degrees. (The physical displacement between poles on the stator frame is $180^\circ/p$ for a machine

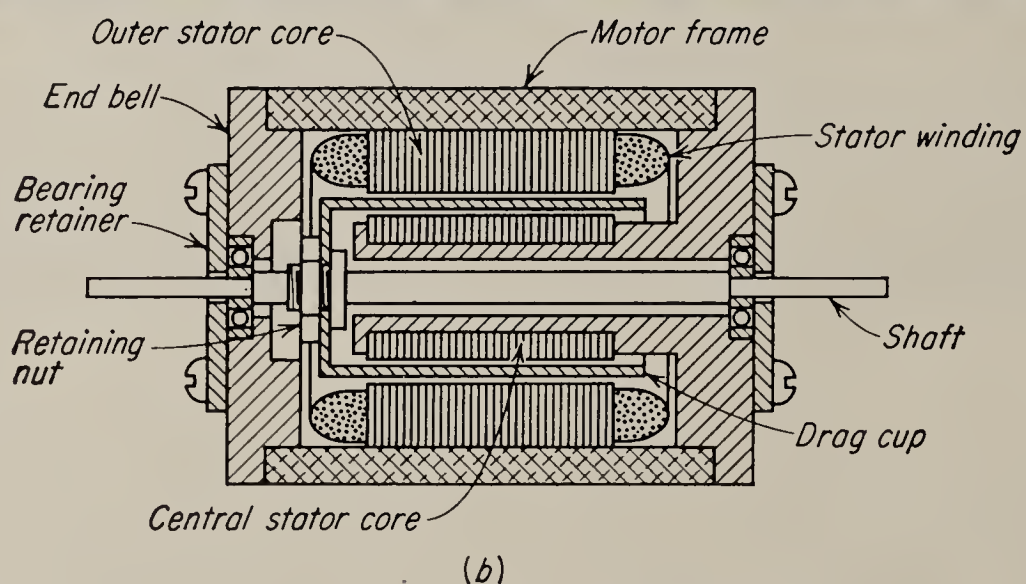
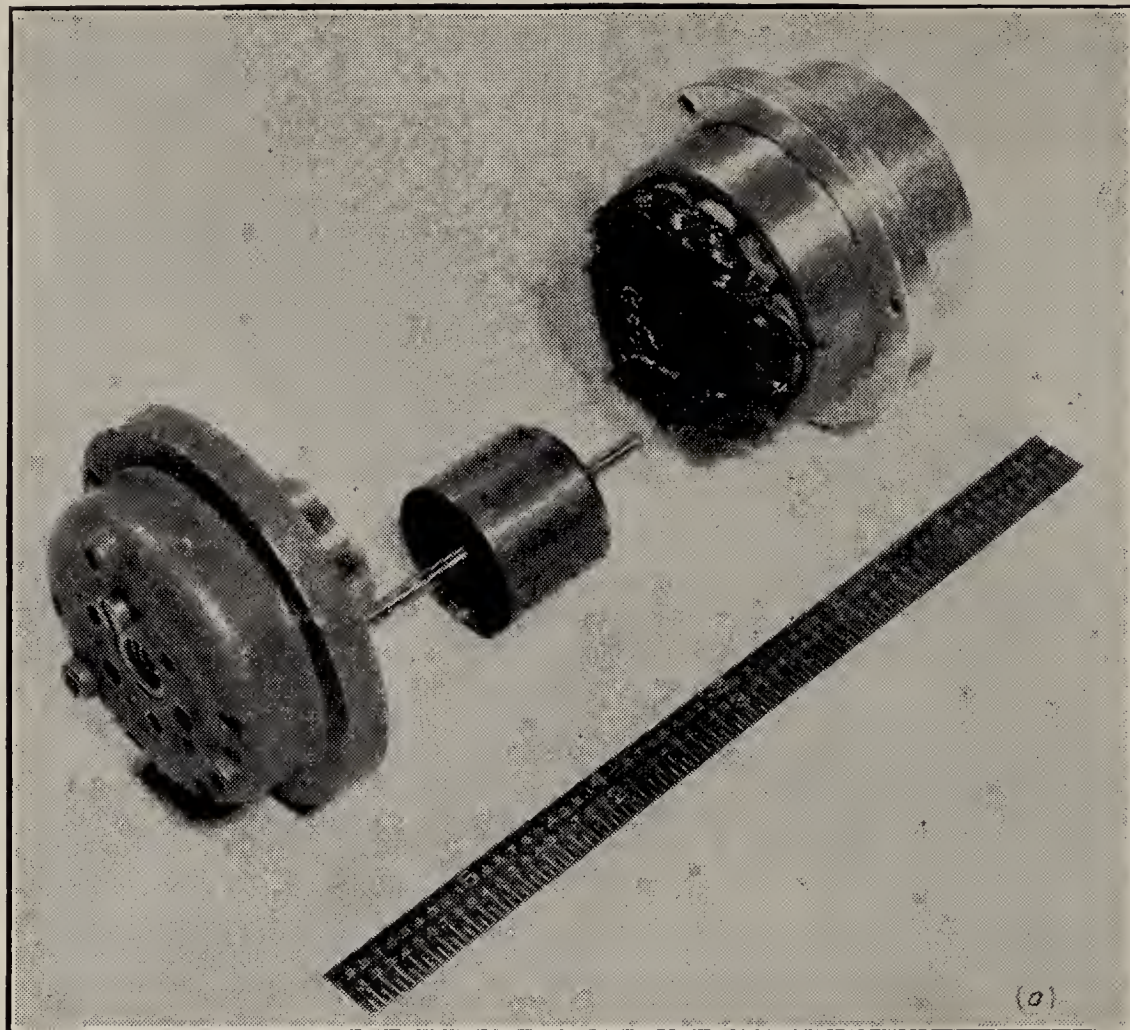


FIG. 7.4. (a) Exploded view of a drag-cup a-c servomotor; (b) internal construction of a drag-cup motor.

wound for p poles.) These windings are excited by an alternating voltage. In the standard polyphase machine the two voltages have the same rms value but are 90° out of phase. The resulting currents set up a rotating magnetic field in the air gap of the machine, and if the windings are properly distributed in the slots, this magnetic field is approximately

uniform in time and a sinusoidal function of angle. The magnetic flux cutting the conductors of the rotor winding set up currents in the rotor, and the interaction of these currents with the flux results in a torque tending to turn the rotor in the same direction as the magnetic field.

A more quantitative insight into motor operation is obtained by the use of an equivalent circuit. To obtain this equivalent circuit, we assume, for the moment, that the rotor is not turning; the rotor may then be thought of as a short-circuited secondary winding of a transformer whose primary is the stator winding. In a two-phase motor there are two such transformers, but if a balanced set of polyphase voltages is applied to the motor, they are identical, and we need to consider only one. The voltage induced in the rotor, or transformer secondary, is proportional to the primary self-induced voltage, the constant of proportionality being the effective turns ratio between primary and secondary. The current per phase flowing in the rotor is therefore

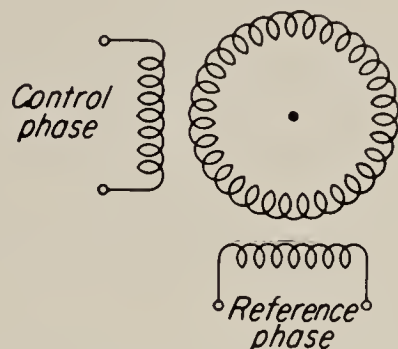


FIG. 7.5. Schematic of two-phase motor.

$$I_{2o} = \frac{E_{2o}}{R_2 + j2\pi fL_2} = \frac{N_2}{N_1} \frac{E_1}{R_2 + j2\pi fL_2} \quad (7.1)$$

In these equations E_1 and E_2 are induced voltages per phase, and N_1 and N_2 are the effective turns of the stator and rotor, respectively. R_2 and L_2 are the resistance and inductance per phase, respectively, of the rotor, and f is the line frequency. The additional subscript o indicates that the quantities so marked are measured at standstill. If the rotor is permitted to rotate, the relative velocity between the rotating magnetic field and the rotor conductors decreases. Hence both the voltage induced in the rotor and the rotor frequency decrease, becoming zero when the rotor speed becomes *synchronous*, i.e., equal to the speed of the magnetic field. It is easily demonstrated that the induced voltage and frequency are proportional to the *slip* S , defined as

$$S = \frac{\text{synchronous speed} - \text{rotor speed}}{\text{synchronous speed}} \quad (7.2)$$

Hence, in general, the rotor current per phase becomes

$$I_2 = \frac{SE_{2o}}{R_2 + j2\pi fSL_2} = \frac{N_2}{N_1} \frac{SE_1}{R_2 + jSX_{2o}} \quad (7.3)$$

where X_{2o} is the rotor reactance per phase at standstill. Equation (7.3) may be written in the equivalent form

$$I_2 = \frac{N_2}{N_1} \frac{E_1}{(R_2/S) + jX_{2o}} = \frac{N_2}{N_1} \frac{E_1}{R_2 + jX_{2o} + R_2[(1 - S)/S]} \quad (7.4)$$

In the second form of Eq. (7.4) the effective rotor resistance R_2/S has been broken up into two components, R_2 and $R_2[(1 - S)/S]$. This is done to indicate two components of rotor power. The component due to the currents flowing through R_2 , that is, $|I_2|^2 R_2$, is the rotor copper loss; hence the remaining power, $|I_2|^2 R_2[(1 - S)/S]$, is proportional to the mechanical power developed by the rotor.

In order to determine the stator current, we note that, whether or not the rotor turns, the points of both maximum induced voltage and maximum rotor current must rotate around the rotor in synchronism with the stator magnetic field. The rotor current therefore causes a demagnetizing mmf $N_2 I_2$ rotating at synchronous speed (note that a squirrel-cage winding is effectively a single turn; hence the mmf is numerically

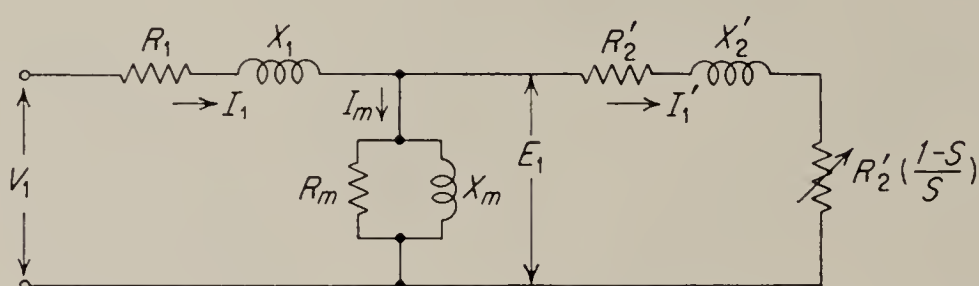


FIG. 7.6. Induction-motor equivalent circuit.

equal to I_2). Since the magnetic flux in the air gap is fixed if the applied voltage is fixed, a current must flow in the stator to produce a magnetizing force exactly equal to the demagnetizing $N_2 I_2$; i.e.,

$$N_1 I_1' = N_2 I_2 \quad (7.5)$$

We use the symbol I_1' , since the current required to cancel the mmf $N_2 I_2$ is only the *power-producing component* of the stator current. There is also a *magnetizing component* I_M , which is required to set up the rotating magnetic field. From Eqs. (7.5) and (7.4), we get

$$I_1' = \frac{E_1}{R_2' + jX_{2o}' + R_2'[(1 - S)/S]} \quad (7.6)$$

where $R_2' = R_2(N_1/N_2)^2$ and $X_{2o}' = X_{2o}(N_1/N_2)^2$. The primary induced voltage E_1 may be obtained in terms of the applied voltage and the stator current by writing

$$E_1 = V_1 - I_1(R_1 + jX_1) \quad (7.7)$$

where V_1 is the applied voltage per phase, $I_1 = I_1' + I_M$ is the stator current, and R_1 and X_1 are the resistance and leakage reactance of the stator, respectively. Equations (7.6) and (7.7) form the basis of the equivalent circuit, which takes the form shown in Fig. 7.6. In this circuit X_M is the magnetizing reactance, and the current flowing in it is the magnetizing component of the primary current. The resistance R_M is added to account for the fact that the magnetizing current contains a small inphase

component to supply the iron losses. In most servomotors R_M is so much larger than X_M that it may often be ignored.

The total power going into the rotor is $|I'_1|^2 R'_2 / S$. Note that

$$|I'_1|^2 R'_2 = |I_2|^2 R_2 = \text{rotor copper loss}$$

so that $|I'_1|^2 R'_2 [(1 - S)/S]$ must therefore be the developed mechanical power.

To find the developed torque, use may be made of the fact that, if a consistent set of units is used, the product of torque and speed is power. Using Q for torque and Ω for speed, we have

$$Q\Omega = 2|I'_1|^2 R'_2 \frac{1 - S}{S}$$

The factor 2 arises from the fact that the analysis has been made on a *per phase* basis. Since there are two phases, the actual power is therefore twice the power per phase. From Eq. (7.2) we see that $1 - S = \Omega/\Omega_s$, where Ω_s is the synchronous speed. Hence

$$Q = \frac{2}{\Omega_s} |I'_1|^2 \frac{R'_2}{S} \quad (7.8)$$

Since Ω_s is a constant, the developed torque is seen to be proportional to the total power dissipated in the rotor.

We can now obtain an expression for the torque as a function of the input voltage V_1 and the slip. To simplify the algebra, we use the notation

$$\begin{aligned} Z_1 &= R_1 + jX_1 \\ \frac{1}{Z_M} &= \frac{1}{R_M} + \frac{1}{jX_M} \\ Z_2(S) &= \frac{R'_2}{S} + jX'_2 \end{aligned}$$

Then
$$I'_1 = \frac{E_1}{Z_2(S)}$$

and
$$E_1 = V_1 \frac{Z_2(S)Z_M/[Z_2(S) + Z_M]}{Z_1 + \{Z_2(S)Z_M/[Z_2(S) + Z_M]\}}$$

Therefore
$$Q = \frac{2}{\Omega_s} \left| \frac{\frac{Z_M}{Z_2(S) + Z_M} V_1}{Z_1 + \frac{Z_2(S)Z_M}{Z_2(S) + Z_M}} \right|^2 \frac{R'_2}{S} = F(S)|V_1|^2 \quad (7.9)$$

7-4. Theory of Operation of A-C Servomotors. The operation of a-c servomotors differs from that of conventional induction motors primarily because the voltages applied to the servomotor usually do not constitute a balanced polyphase set. Hence we must investigate the behavior of induction motors supplied by unbalanced voltages.

In the following¹ analysis we assume that the two stator windings of the servomotor are identical and that both the reference voltage and the control voltage are supplied from sources having zero output impedance. Extensions of the theory to cases in which these assumptions do not hold

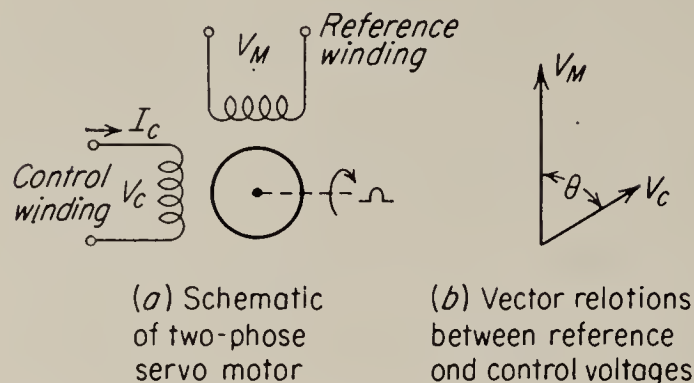


FIG. 7.7. Nomenclature to be used in analysis of two-phase servomotor.

are not difficult and are indicated below. The nomenclature to be used is indicated in Fig. 7.7. The reference voltage V_M is assumed to be constant; the control voltage V_c is variable. The vector diagram shows the general relation existing between V_M and V_c . The angle θ is normally 90° , but for the sake of generality it is taken to have any arbitrary fixed value.

Since the analysis of polyphase motors with balanced voltages applied to the windings is relatively simple, it is convenient to use the method of symmetrical components to convert the unbalanced set of voltages shown in Fig. 7.7 into two balanced sets of voltages of opposite phase rotation.² The two symmetrical components are shown in Fig. 7.8 for a particular set of values of V_M , V_c , and θ . Note that V_{c1} lags V_{M1} by 90° , and V_{c2} leads V_{M2} by 90° . The magnitude of the two symmetrical components

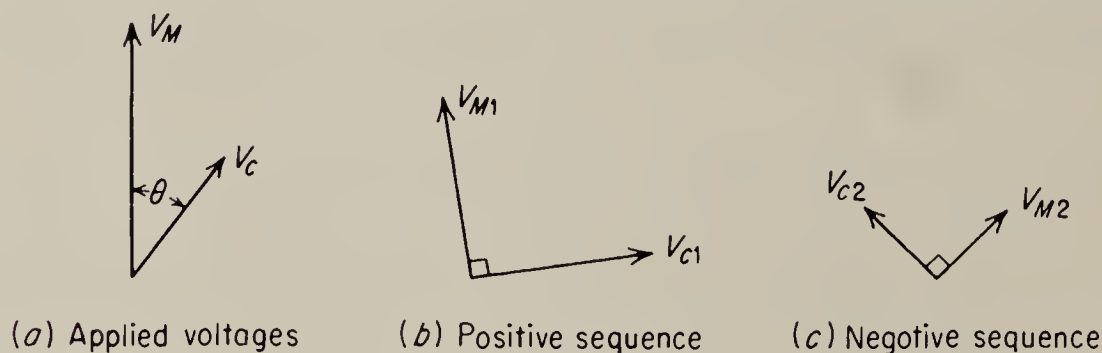


FIG. 7.8. Voltages applied to motor and their symmetrical components.

and their relative phase position depends on the magnitudes of V_M and V_c and θ .

To make the transformation we write

$$V_c = V_{c1} + V_{c2} \quad (7.10)$$

$$V_M = V_{M1} + V_{M2} \quad (7.11)$$

But $V_{c1} = -jV_{M1} \quad (7.12)$

$$V_{c2} = jV_{M2} \quad (7.13)$$

so that $V_c = -jV_{M1} + jV_{M2} \quad (7.14)$

$$jV_c = V_{M1} - V_{M2} \quad (7.15)$$

¹ This section essentially follows Koopman, Operating Characteristics of 2-phase Servomotors, *Trans. AIEE*, vol. 68, pp. 319-329, 1949.

² Lyon, "Application of the Method of Symmetrical Components," McGraw-Hill Book Company, Inc., New York, 1937, pp. 96-110.

Then a simultaneous solution of Eqs. (7.11) and (7.15) gives

$$V_{M1} = \frac{V_M + jV_c}{2} \quad (7.16)$$

$$V_{M2} = \frac{V_M - jV_c}{2} \quad (7.17)$$

Now suppose that the ratio of V_c to V_M is k and that V_c lags V_M by an angle θ . Then

$$V_c = kV_M / -\theta \quad (7.18)$$

and therefore

$$V_{M1} = \frac{V_M}{2} (1 + k/90^\circ - \theta) = \frac{V_M}{2} (1 + k \sin \theta + jk \cos \theta) \quad (7.19)$$

$$V_{M2} = \frac{V_M}{2} (1 - k/90^\circ - \theta) = \frac{V_M}{2} (1 - k \sin \theta - jk \cos \theta) \quad (7.20)$$

so that $|V_{M1}|^2 = \frac{V_M^2}{4} (1 + 2k \sin \theta + k^2) \quad (7.21)$

$$|V_{M2}|^2 = \frac{V_M^2}{4} (1 - 2k \sin \theta + k^2) \quad (7.22)$$

When the unbalanced set of voltages is applied to the motor, the theory of symmetrical components indicates that the motor acts as though both symmetrical sets of voltage were applied simultaneously. Hence, if we assume that the electric and magnetic circuits making up the motor are linear, we may say that the torque developed by the motor is equal to the difference between the torque produced by the positive-sequence voltage and that produced by the negative sequence.

Equation (7.9) gives an expression for the torque produced by a balanced set of voltages; suppose this set to be the positive sequence. The expression for torque produced by the negative sequence must have the same form as (7.9) except that, if the slip is to retain its zero reference at *positive* synchronous speed, $(2 - S)$ must be substituted for S . Since Eq. (7.9) shows the torque to be proportional to the square of the applied voltage we find that in general

$$Q = F(S)|V_{M1}|^2 - F(2 - S)|V_{M2}|^2$$

or, if we let $|V_M|^2 F(S) = Q_b(S)$, the torque produced with balanced input voltages, then is

$$Q = Q_b(S) \left| \frac{V_{M1}}{V_M} \right|^2 - Q_b(2 - S) \left| \frac{V_{M2}}{V_M} \right|^2 \quad (7.23)$$

If we substitute for V_{M1} and V_{M2} their values as obtained in Eqs. (7.21) and (7.22), we find that

$$Q = \frac{1}{4}[Q_b(S)(1 + 2k \sin \theta + k^2) - Q_b(2 - S)(1 - 2k \sin \theta + k^2)] \quad (7.24)$$

Equation (7.24) indicates the motor to be generally nonlinear, since in addition to the squared term in k , $Q_b(S)$ is not a linear function of S . It can, however, be shown that the starting torque of the motor is a linear function of the applied voltage. When the motor is not running, $S = 1$, $Q_b(S) = Q_b(2 - S) = Q_b(1) = Q_{b0}$, and therefore

$$Q_{\text{starting}} = Q_{b0}k \sin \theta \quad (7.25)$$

The presence of the $\sin \theta$ term indicates that the developed stall torque is a function of the phase angle between the control and reference voltages, so that the motor may be thought of as a phase discriminator. Normally, of course, θ is fixed at 90° , since this results in maximum torque. Equation (7.25) also implies that the motor is approximately linear as long as the speed is low, i.e., as long as $Q_b(S) \approx Q_b(2 - S)$.

Equation (7.24) may be used to plot a complete set of speed-torque curves for the motor, provided that $Q_b(S)$ is known for all S of interest. This function is defined in terms of the impedances of the motor as in Eq. (7.9) and can, therefore, be computed if these impedances are known. The measurement of the impedances is, however, always rather difficult, and in particular it is almost impossible to separate rotor and stator impedances by means of simple measurements. In large motors the approximation is therefore made that Z_M is very large, so that its effect is negligible; then we see from Eq. (7.9) that $Q_b(S)$ becomes approximately

$$Q_b(S) \approx \frac{2}{\Omega_s} \frac{|V_M|^2 R_2 S}{|R_2 + SR_1 + jS(X_1 + X_2)|^2}$$

If this approximation is valid, then all that needs to be measured is R_1 , R_2 , and $X_1 + X_2$. The stator resistance R_1 can be measured by a d-c test; $R_1 + R_2$ and $X_1 + X_2$ are obtainable from a measurement of the motor input impedance at standstill ($S = 1$); and if the further assumption is then made that R_2 is independent of speed, $Q_b(S)$ can be computed for any desired value of S .

In small servomotors, particularly in those with a drag-cup rotor, the assumption that Z_M is very large is not particularly valid. Furthermore, R_2 depends to some extent on the rotor frequency because of skin effect and eddy currents. Hence a computation of $Q_b(S)$, in addition to being rather laborious, will not give very accurate results. The problem is usually, therefore, completely bypassed by measuring the speed-torque curve of the motor with balanced input, that is, $Q_b(S)$, rather than the impedances. This is quite easy to do for speeds between zero and synchronous, and this type of curve is usually furnished by the motor manufacturer as part of the motor characteristics. The only difficulty is that $Q_b(S)$ must be known for values of S ranging between 0 and 2; i.e., a speed-torque curve for speeds from synchronous forward to synchronous

backward is necessary. Since the speed-torque curve for negative speed requires special measuring equipment and is usually not supplied by the manufacturer, it becomes of interest to be able to determine $Q_b(S)$ for all S given its value for $0 \leq S \leq 1$. This may be done as follows:

By means of a simple algebraic manipulation, $Q_b(S)$ as given in Eq. (7.9) may be written in the equivalent form

$$Q_b(S) = 2 \left| \frac{Z_M V_M}{Z_1 + Z_M} \right|^2 \left| \frac{1}{R'_2 + jSX'_2 + S[Z_1 Z_M / (Z_1 + Z_M)]} \right|^2 \frac{R'_2 S}{\Omega_s} \quad (7.26)$$

The term $Z_1 Z_M / (Z_1 + Z_M)$ is an equivalent impedance, which may be written as $R_0 + jX_0$. Since Z_M is usually much larger than Z_1 , $|R_0 + jX_0|$ is approximately equal to $|R_1 + jX_1|$, but since $Z_M \approx jX_M$, the phase angle of $R_0 + jX_0$ is always greater than that of $R_1 + jX_1$. Hence R_0 is usually a very small resistance. Also in most servomotors, R'_2 is quite large. Hence

$$\left| R'_2 + jSX'_2 + S \frac{Z_1 Z_M}{Z_1 + Z_M} \right|^2 = |R'_2 + SR_0 + jS(X'_2 + X_0)|^2 \approx R'^2_2 + S^2(X'_2 + X_0)^2 \quad (7.27)$$

The term $|Z_M V_M / Z_1 + Z_M|^2$ is a constant, independent of S ; hence $Q_b(S)$ has the general form

$$Q_b(S) = \frac{S}{C_1 + C_2 S^2} \quad (7.28)$$

In writing Eq. (7.28), we make the further assumption that R'_2 is independent of S . Strictly speaking, this is not true, since skin effect causes R_2 to increase with frequency and, therefore, with S . This effect should be particularly noticeable in motors designed for operation at 400 cps and having either solid-iron rotors or squirrel-cage rotors with very deep rotor bars. The skin effect is relatively negligible in lower-frequency motors, such as 60 cps, and in motors having drag-cup rotors. At any rate, it is clear that Eq. (7.28) is an approximation, but it is a very simple equation and sufficiently accurate in most cases. Since there are only two constants, C_1 and C_2 , to be evaluated, it is only necessary to know the value of $Q_b(S)$ for two values of S other than zero to be able to compute $Q_b(S)$ for all other values of S . This means, of course, that the knowledge of $Q_b(S)$ for $0 \leq S \leq 1$ is quite sufficient, if extreme accuracy is not required.

A typical set of motor characteristics is shown in Fig. 7.9; the curve for $k = 1$ was taken from the data supplied by the manufacturer for $0 \leq S \leq 1$; Eq. (7.28) was used to extend this curve into the region $1 \leq S \leq 2$ by using the given data for $S = 1$ and $S = 1/2$, and Eq. (7.24) was then used to compute the curves for other values of k . The phase angle between reference and control voltage was taken as 90° . Note that near zero speed the curves are very nearly parallel and straight.

7.5. Approximate Transfer Function of the A-C Servomotor. Strictly speaking, the term “transfer function” has meaning only for linear systems; however an approximate transfer function may often be defined for a nonlinear device, such as the a-c servomotor, if an operating point is specified and if the range of input and output variables is small enough to

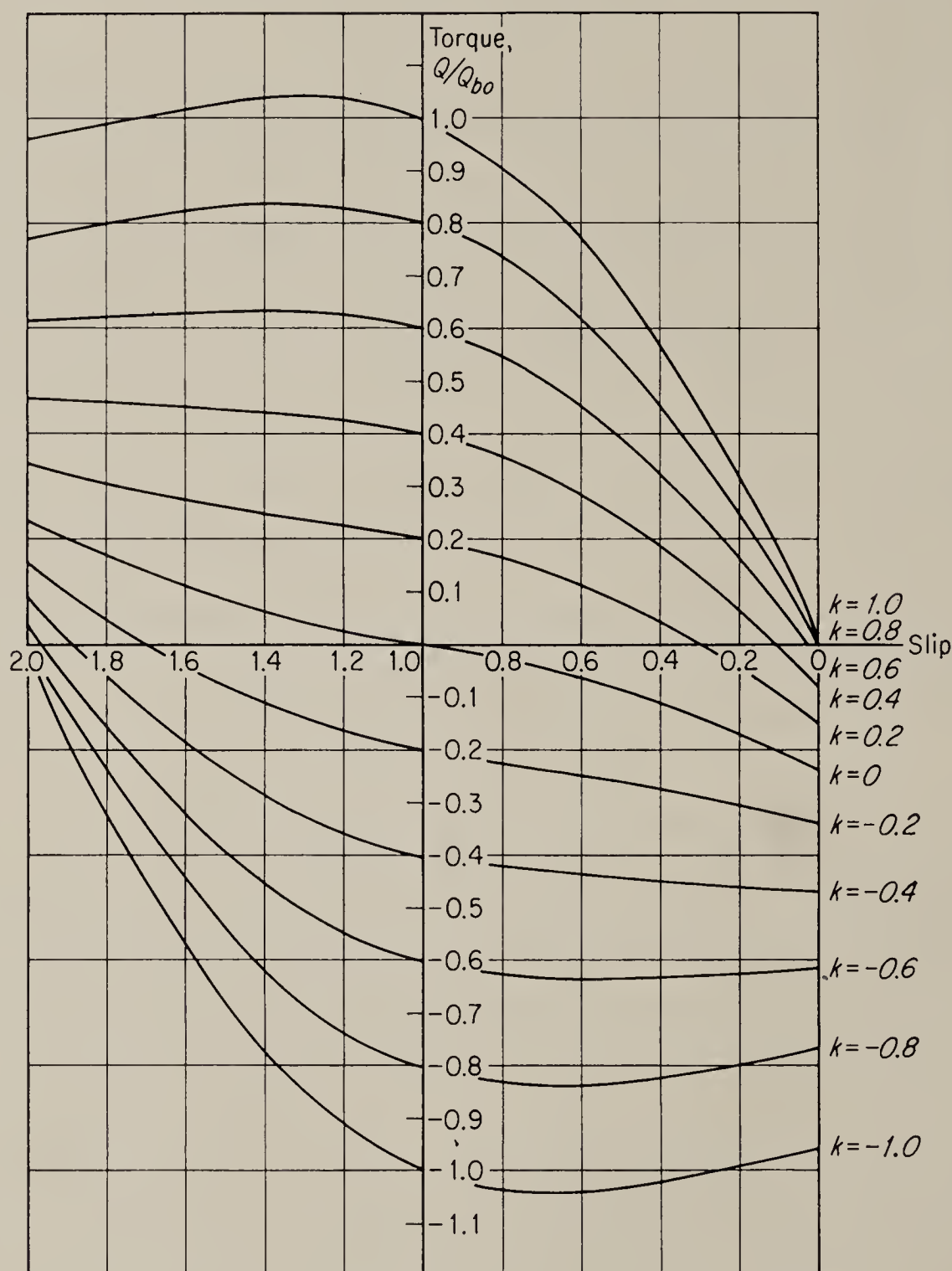


FIG. 7.9. Typical speed-torque curves of a-c servomotor.

permit the characteristics to be approximated to a sufficient degree of accuracy by the first-order term of the Taylor expansion. In order to obtain such a transfer function for the a-c servomotor, we first express its speed-torque characteristics in the form

$$\Delta\Omega = A \Delta V_c - B \Delta Q \quad (7.29)$$

where $A = \partial\Omega/\partial V_c$ and $B = -\partial\Omega/\partial Q$, evaluated at the desired operating point, and $\Delta\Omega$, ΔV_c , and ΔQ are small variations of the speed, applied con-

trol voltage, and developed torque, respectively, from their operating-point values. By analogy with the d-c motor (see Chap. 4), A will be recognized as the motor gain constant, and $1/B$ as the effective coefficient of viscous damping.

Given Eq. (7.29), a transfer function can be found. The small electrical time lag between the application of the control voltage and the development of torque is neglected for the moment, and the mechanical friction is assumed negligible. Then the developed torque is opposed solely by the torque required to accelerate the moment of inertia and by a possible arbitrary load torque Q_a . If the moment of inertia is J , then, in Laplace transform notation,

$$\begin{aligned}\hat{\Omega} &= A\hat{V}_c - B(Js\hat{\Omega} + \hat{Q}_a) \\ \text{or} \quad \hat{\Omega} &= \frac{A\hat{V}_c - B\hat{Q}_a}{JBs + 1}\end{aligned}\tag{7.30}$$

In these equations the symbol $\hat{\Omega}$ stands for the transform of $\Delta\Omega$, etc. If the small electrical time lag due to the inductance of the stator and rotor circuits is not negligible, then, approximately,

$$\hat{\Omega} = \frac{A\hat{V}_c - B\hat{Q}_a}{(T_1s + 1)(JBs + 1)}\tag{7.31}$$

where T_1 is the stator-inductance time constant neglected in Eq. (7.30); it is typically about ten times smaller than the inertia time constant JB .*

To evaluate A and B from the speed-torque curves, we note first that, since

$$V_c = kV_M$$

$$\text{and} \quad \Omega = \Omega_s(1 - S)$$

$$\text{therefore} \quad \frac{\partial \Omega}{\partial V_c} = -\frac{\Omega_s}{V_M} \frac{\partial S}{\partial k}$$

$$\text{and} \quad \frac{\partial \Omega}{\partial Q} = -\Omega_s \frac{\partial S}{\partial Q}$$

Both $\partial S/\partial k$ and $\partial S/\partial Q$ may be found by partial differentiation of Eq. (7.24). Thus

$$\begin{aligned}A &= \frac{\partial \Omega}{\partial V_c} \\ &= \frac{2\Omega_s}{V_M} \frac{(\sin \theta + k)Q_b(S) + (\sin \theta - k)Q_b(2 - S)}{\frac{\partial Q_b(S)}{\partial S} (1 + 2k \sin \theta + k^2) - \frac{\partial Q_b(2 - S)}{\partial S} (1 - 2k \sin \theta + k^2)}\end{aligned}\tag{7.32}$$

* L. O. Brown, Jr., Transfer Function for a Two-phase Induction Servo Motor, *Trans. AIEE*, vol. 70, pp. 1890-1893, 1951.

and

$$B = -\frac{\partial \Omega}{\partial Q} = \frac{4\Omega_s}{\frac{\partial Q_b(S)}{\partial S}(1 + 2k \sin \theta + k^2) - \frac{\partial Q_b(2 - S)}{\partial S}(1 - 2k \sin \theta + k^2)} \quad (7.33)$$

As an example of the application of Eqs. (7.32) and (7.33), let us use the simplified form of the expression for $Q_b(S)$ developed in the previous section [see Eq. (7.28)],

$$Q_b(S) = \frac{S}{C_1 + C_2 S^2} \quad (7.34)$$

and use this to find A and B at the operating point ordinarily of greatest interest in servo applications, i.e., the point at which $S = 1$, $k = 0$. We obtain

$$A = 2 \frac{\Omega_s}{V_M} \frac{C_1 + C_2}{C_1 - C_2} \sin \theta$$

$$\text{and} \quad B = 2\Omega_s \frac{(C_1 + C_2)^2}{C_1 - C_2} \quad (7.35)$$

In many servomotors, $Q_b(S)$ is so close to being a linear function of S that C_2 in Eq. (7.34) becomes negligible. Under these conditions, $1/C_1$ becomes $Q_b(1) = Q_{b0}$, the blocked-rotor torque, and thus

$$A = 2 \frac{\Omega_s}{V_M} \sin \theta \quad (7.36)$$

$$B = 2 \frac{\Omega_s}{Q_{b0}} \quad (7.37)$$

If, on the other hand, C_2 is larger than C_1 , both A and B become negative. This represents a form of unstable behavior referred to as *single-phasing*, which is discussed in more detail in a later section. By differentiating Eq. (7.34) it can be shown that, when C_2 is greater than C_1 , $Q_b(S)$ peaks for $1 > S > 0$, i.e., for positive speeds.

The process of partial differentiation described above is relatively rigorous, but it is rather involved. Hence a simpler, although less accurate, method is often employed to obtain A and B . In this method the speed-torque curves of the motor are replaced by parallel straight lines, as shown in Fig. 7.10. The line for $k = 1$, $\theta = 90^\circ$ is drawn through the points $\Omega = \Omega_s$, $Q = 0$ and $\Omega = 0$, $Q = Q_{b0}$, and the lines for various k 's and θ 's are drawn parallel to the first one and such that the blocked-rotor torque is proportional to $k \sin \theta$. It is clear that, since this method considers only the end points of the actual speed-torque curve of the motor, it cannot take into account any curvature in this characteristic. Also it is apparent that it does not yield the same answer as the partial-differ-

entiation method, even when $Q_b(S)$ is assumed to be linear. For from the linear characteristics of Fig. 7.10 we get

$$A = \frac{\Omega_s}{V_M} \sin \theta \quad (7.38)$$

$$B = \frac{\Omega_s}{Q_{b0}} \quad (7.39)$$

which is half as much as was obtained by the partial-differentiation method [Eqs. (7.36) and (7.37)]. However, since the method yields the gain and time constant almost directly by inspection, with the only data needed being the blocked torque and synchronous speed, it is very convenient when only a rough estimate of the motor transfer function is required.

7.6. Figures of Merit for A-C Motors.

At this point it is perhaps appropriate to discuss two figures of merit for servomotors that are commonly used: the *torque-to-inertia ratio* and the *torque-squared-to-inertia ratio*. In both figures of

merit the torque in question is Q_{b0} , the stalled torque obtained when balanced polyphase voltages are applied to the motor. The torque-to-inertia ratio gives an indication of the acceleration capabilities of the motor without load and can be shown to be equivalent to the product of bandwidth and maximum speed, which, in turn, is somewhat analogous to the voltage-gain-bandwidth product often used as a figure of merit in amplifiers. For, if the bandwidth is defined as the reciprocal of the motor-inertia time constant, the product of speed and bandwidth is Ω_s/JB , and if the simplified definition of B given in Eq. (7.39) is used, this becomes Q_{b0}/J , the torque-to-inertia ratio. The torque-squared-to-inertia ratio can be shown to be equivalent to the product of bandwidth and power output, which is a criterion similar to the power-gain-bandwidth product used for power amplifiers. If the motor speed-torque curve is a straight line, the torque is given by

$$Q = Q_{b0} \left(1 - \frac{\Omega}{\Omega_s} \right)$$

and therefore the developed power is

$$Q\Omega = Q_{b0} \left(\Omega - \frac{\Omega^2}{\Omega_s} \right)$$

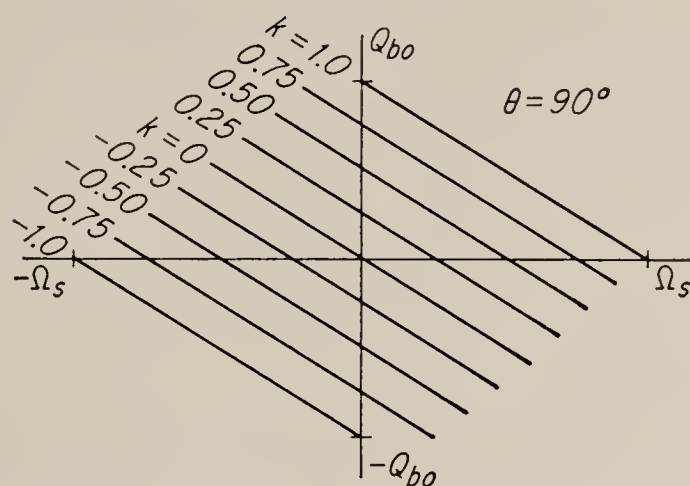


FIG. 7.10. Linearized speed-torque curves for a-c servomotor.

The maximum power is found by differentiation to be equal to $Q_{b0}\Omega_s/4$, and therefore the power-bandwidth product becomes $1/4 Q_{b0}\Omega_s/JB$. If the simplified definition for B from Eq. (7.39) is again used, this becomes $1/4 Q_{b0}^2/J$. The power-bandwidth product is a somewhat more universal criterion than the speed-bandwidth product and is not affected by gearing interposed between motor and load. Hence, when the gearing between motor and load is such that the load inertia has an appreciable effect on the motor, the power-bandwidth product or torque-squared-to-inertia ratio gives a somewhat better indication of motor merit than the torque-to-inertia ratio.

The torque-to-inertia ratio is increased by decreasing the diameter of the rotor. This is due to the fact that for constant rotor length, the inertia is proportional to the fourth power of the diameter, whereas the torque decreases only at something between the second and third power of the diameter. The torque-squared-to-inertia ratio is seen by the same reasoning to be decreased by a decrease of the diameter; hence an optimum ratio of diameter to length exists, its value depending on how the motor is to be used.

Torque-to-inertia ratios of motors available commercially range from about 40 to 120 in.⁻¹ with the ratio decreasing for motors of larger power rating. Motors designed for 400 cps generally have a somewhat lower figure than 60-cps motors. Torque-squared-to-inertia ratios vary from about 10 to over 60 lb. It should be pointed out that, although figures of merit are convenient means for comparing motors, no single figure of this sort can be expected to express the merit of the motor adequately in all situations. Thus, figures such as ratio of stall torque to watt input, etc., are sometimes more important, and, of course, in commercial applications cost is a very important consideration and may often dictate the use of a motor whose characteristics, as measured by the figures of merit described here, are quite inferior.

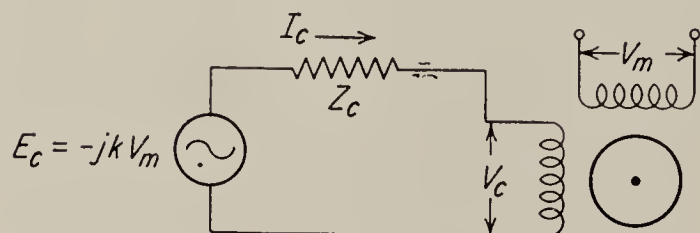


FIG. 7.11. Servomotor controlled by source having finite output impedance.

7.7. Motor Characteristics in the Presence of Finite Control Impedance. In the development carried out thus far it has been assumed that the source of control voltage has zero output impedance. In general, when a motor is controlled

by an electronic amplifier, the output impedance is not zero, and the control voltage V_c is then a function of the current I_c in the control winding.

To extend the theory to embrace this possibility, let it be supposed that the internal voltage of the amplifier is $E_c = jkV_M$ and that the amplifier has an output impedance Z_c (see Fig. 7.11). In order to keep the discussion as simple as possible, we shall assume that E_c and V_M are 90° out of

phase, although in principle there is no difficulty in considering other phase angles between the two voltages. Since the current in the windings plays a role under the conditions assumed here, it will be found convenient to consider the symmetrical components of the current as well as those of the voltage. Thus, let

$$I_c = I_{c1} + I_{c2} \quad (7.40)$$

The input impedance per phase presented by the motor to the positive sequence is essentially the impedance of the equivalent circuit of Fig. 7.6. This impedance is

$$Z = Z_1 + \frac{Z_M[(R'_2/S + jX'_2)]}{Z_M + (R'_2/S + jX'_2)} \triangleq Z(S) \quad (7.41)$$

where $Z_1 = R_1 + jX_1$ and $Z_M = \frac{jR_M X_M}{R_M + jX_M}$

Then the impedance presented to the negative sequence is $Z(2 - S)$. Hence

$$I_{c1} = \frac{V_{c1}}{Z(S)} \quad (7.42)$$

and
$$I_{c2} = \frac{V_{c2}}{Z(2 - S)} \quad (7.43)$$

Also, from the circuit of Fig. 7.11,

$$V_c = E_c - I_c Z_c = jkV_M - I_c Z_c \quad (7.44)$$

From previous developments [Eqs. (7.12) to (7.14)] we have

$$\begin{aligned} V_{c1} &= -jV_{M1} \\ V_{c2} &= jV_{M2} \\ V_c &= V_{c1} + V_{c2} = -j(V_{M1} - V_{M2}) \end{aligned}$$

Hence Eq. (7.44) becomes

$$-kV_M - Z_c \left[\frac{V_{M2}}{Z(2 - S)} - \frac{V_{M1}}{Z(S)} \right] = V_{M2} - V_{M1} \quad (7.45)$$

From Eq. (7.11)

$$V_{M2} = V_M - V_{M1}$$

so that finally

$$V_{M1} = V_M \frac{1 + k + [Z_c/Z(2 - S)]}{2 + [Z_c/Z(S)] + [Z_c/Z(2 - S)]} \quad (7.46)$$

and similarly

$$V_{M2} = V_M \frac{1 - k + [Z_c/Z(S)]}{2 + [Z_c/Z(S)] + [Z_c/Z(2 - S)]} \quad (7.47)$$

The developed torque can now be found from Eq. (7.23) and is given by the expression

$$Q = \frac{1}{\left| 2 + \frac{Z_c}{Z(S)} + \frac{Z_c}{Z(2-S)} \right|^2} \left[Q_b(S) \left| 1 + k + \frac{Z_c}{Z(2-S)} \right|^2 - Q_b(2-S) \left| 1 - k + \frac{Z_c}{Z(S)} \right|^2 \right] \quad (7.48)$$

It is apparent that the presence of Z_c results in a rather complicated expression, especially since it is now necessary to know the input impedance of the motor in addition to the speed-torque curve for balanced voltages. The remarks made concerning the difficulty of computing $Q_b(S)$ apply with even greater force to the computation of $Z(S)$, since it is necessary to find not only the magnitude but also the phase angle as a function of S . This information is not ordinarily supplied by the motor manufacturers. Thus, unless a number of rather drastic simplifying assumptions are made or unless $Z(S)$ is measured directly, Eq. (7.48) is difficult to use directly. Figure 7.12 gives a typical set of speed-torque curves for a motor assumed to have the very much simplified equivalent circuit shown; note that the curves for $Z_c = 200 + j0$ are considerably less regular than those for $Z_c = 0$.*

It is interesting to note that the addition of Z_c does not change the fact established previously that the blocked-rotor torque is a linear function of k . If in Eq. (7.48) we let $S = 1$, we obtain, after some simplification,

$$Q = kQ_{b0} \frac{1 + R}{|1 + (Z_c/Z)|^2} \quad (7.49)$$

where R is the real part of Z_c/Z . For speeds close to zero the approximation holds that $Z(S) \approx Z(2-S) \approx Z(1)$, so that Eq. (7.48) can be simplified somewhat. The transfer function of the motor, although basically unchanged, is, of course, also more difficult to find. In particular, the determination of A and B by the "exact" method becomes considerably more cumbersome. It is, however, apparent from the appearance of the curves shown in Fig. 7.12 that both the gain and damping for the operating point $k = 0$, $S = 1$ will be reduced by the presence of Z_c in the control circuit.

7.8. Unbalanced Stator Windings. Quite commonly, motors are built with more turns on the control winding than on the reference winding. The advantage of this construction is that the voltage rating of the control winding may be made high enough to make it possible to connect the motor directly to the output tubes of an electronic power amplifier with-

* The equivalent circuit and the curves are adapted from Koopman's paper referred to earlier.

out an output transformer. The analysis of such a motor follows directly from the analysis of the motor with equal windings.

Suppose that the ratio of reference current to control current required to set up a uniform rotating magnetic field in the motor is N . N will be equal to the turns ratio of the windings if the windings differ only in the number of turns and are in all other respects the same. In any case, N

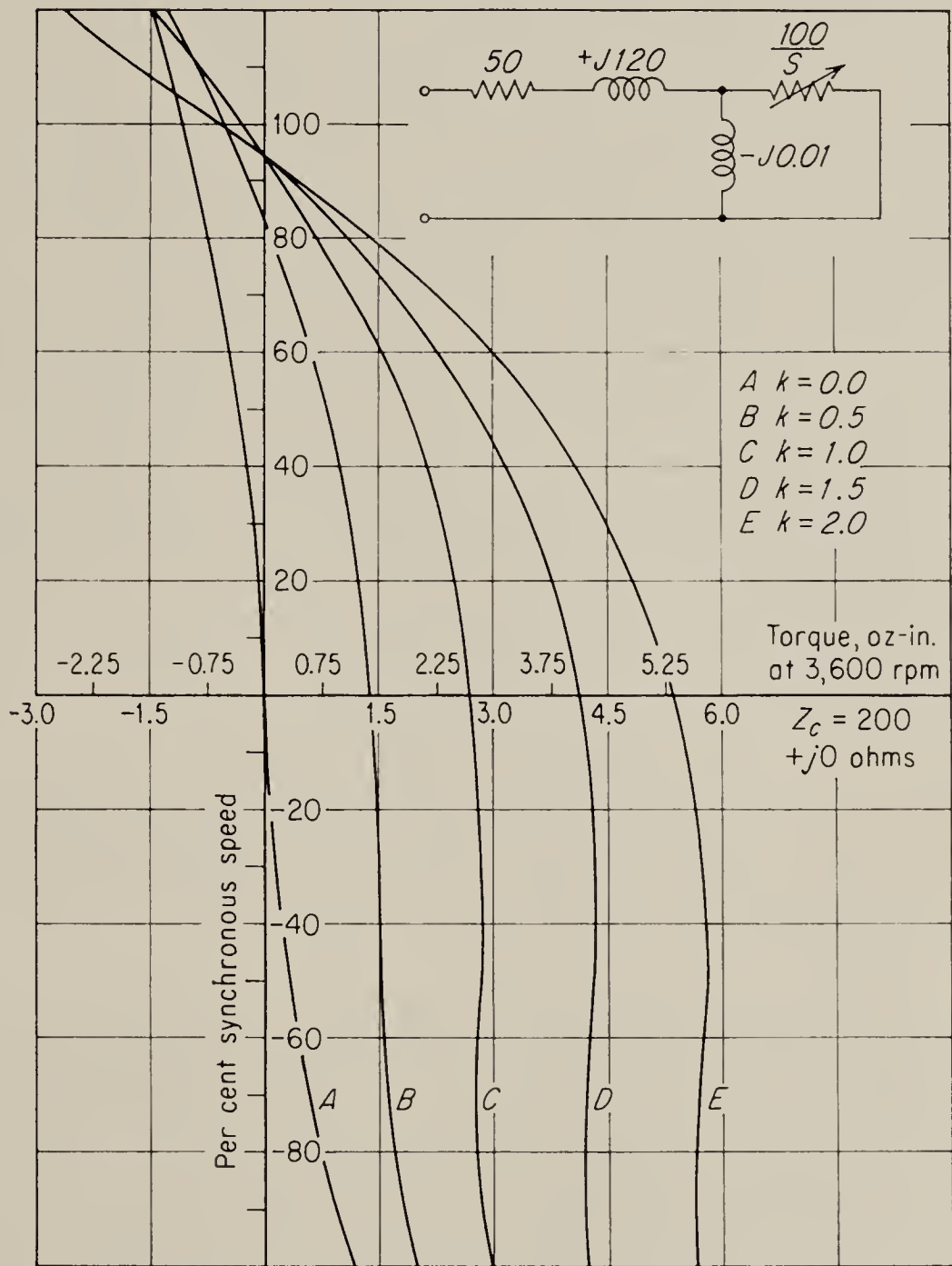


FIG. 7.12. Speed-torque curves for simplified equivalent circuit. (From Koopman)

may be thought of as an equivalent turns ratio. If now the ratio of input impedance of the control winding at any given speed to input impedance of the reference winding for the same speed is equal to N^2 for all S , then the motor behaves exactly as though it had equal windings, but with a transformer of turns ratio N connected between the control winding and the control-voltage source. Such a motor may, therefore, be treated like a motor with equal windings with the equivalent control voltage V_c/N applied to it. If the ratio of control and reference impedance differs from N^2 , we may consider the difference due to a series impedance (which may

turn out to be negative) in the control winding, and the motor may then be analyzed as if it were a motor with equal windings but with an equivalent Z_c in the control winding.

7.9. Single-phasing of an Induction Motor. *Single-phasing* refers to operation of a polyphase induction motor on a single-phase supply. A motor will not single-phase if, when driven by an external torque, it always develops an opposing torque, i.e., if it has a positive damping coefficient. Conversely, a motor that develops torque in the direction of rotation may single-phase if the developed torque exceeds the friction torque; such a motor exhibits negative damping. Although it is theoretically possible to stabilize a servo with a motor having negative damping, it is difficult and requires relatively complicated control circuits. Single-phasing is therefore considered undesirable in a servomotor. Since an ordinary induction motor will run on single-phase, once brought up to speed, it is necessary to incorporate special design features in motors intended for servo use to prevent them from single-phasing. These features are discussed in the following paragraphs. Two cases are considered: (1) the control phase is short-circuited, and (2) the control phase is open-circuited.

If the control phase is short-circuited, $V_c = 0$; hence $k = 0$, and by Eq. (7.24) the torque developed by the motor becomes

$$Q = \frac{1}{4}[Q_b(S) - Q_b(2 - S)] \quad (7.50)$$

If the motor is not to single-phase, it is necessary that the torque be negative when the speed is positive, and vice versa. Since for positive speed $0 \leq S < 1$, we require

$$Q_b(S) - Q_b(2 - S) < 0 \quad \text{for} \quad 0 \leq S < 1 \quad (7.51)$$

But if $S < 1$, $2 - S > 1$; hence the requirement is met if the speed-torque curve for balanced input is such that the torque for all negative speeds exceeds the torque for corresponding positive speeds. In other words, the peak in the speed-torque curve must occur for negative speeds. Figure 7.13a shows such a speed-torque curve and the corresponding speed-torque curve for $k = 0$; in Fig. 7.13b are shown the speed-torque curves for a standard induction motor to illustrate the difference. The speed-torque curve in Fig. 7.13a is characteristic of a motor with a high rotor resistance. This may be demonstrated by differentiating Eq. (7.26) to find the value of S for which $Q_b(S)$ has a peak. It is found that at the peak

$$S^2 = \frac{(R'_2)^2}{R_0^2 + (X'_2 + X_0)^2} \quad (7.52)$$

so that, if the peak is to occur for $S > 1$,

$$(R'_2)^2 > R_0^2 + (X'_2 + X_0)^2 \quad (7.53)$$

The large rotor resistance required to prevent single-phase operation is one of the main reasons for the low efficiency of servomotors.

When the control winding of the motor is open-circuited, the voltage V_c is not, in general, equal to zero, but $Z_c = \infty$. Hence, using Eqs. (7.46) and (7.47), we obtain

$$V_{M1} = V_M \frac{Z(S)}{Z(S) + Z(2 - S)} \quad (7.54)$$

$$\text{and} \quad V_{M2} = V_M \frac{Z(2 - S)}{Z(S) + Z(2 - S)} \quad (7.55)$$

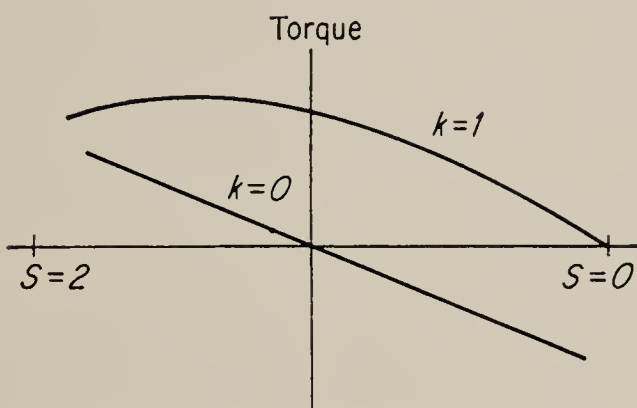
and therefore, by Eq. (7.23),

$$Q = \frac{1}{|Z(S) + Z(2 - S)|^2} [Q_b(S)|Z(S)|^2 - Q_b(2 - S)|Z(2 - S)|^2] \quad (7.56)$$

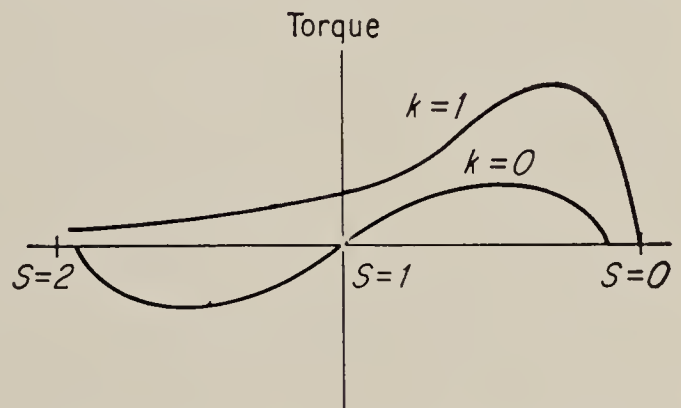
The requirement here becomes, therefore, that

$$Q_b(S)|Z(S)|^2 < Q_b(2 - S)|Z(2 - S)|^2 \quad 0 \leq S < 1 \quad (7.57)$$

if the motor is not to run on single-phase. $Z(S)$ is defined in Eq. (7.41) and increases as S decreases. Hence, even though $Q_b(S) < Q_b(2 - S)$,



(a) Speed-torque curves of servo motor showing positive damping



(b) Speed-torque curves of standard induction motor showing negative damping

FIG. 7.13. Speed-torque curves of a-c motors.

inequality (7.57) is not necessarily satisfied. A motor that will not single-phase when the control phase is short-circuited may therefore single-phase when it is open-circuited. The tendency to single-phase is reduced by making $Q_b(S)$ as much smaller than $Q_b(2 - S)$ as possible, i.e., by using a large rotor resistance. Another technique is to reduce the dependence of $Z(S)$ on S . By inspection of the equivalent circuit (Fig. 7.6) it is seen that a relatively small Z_M and a relatively large R_2 will have this effect. In other words, if most of the current going into the motor is magnetizing current, changes in the load component of input current due to changes in speed will be relatively negligible. Since a large R_2 also serves to make the speed-torque curves more nearly linear, it is found that most servomotors, particularly the smaller ones, where the power

loss is of no consequence, employ rotors with large resistances, and the magnetizing impedance is usually considerably less than in conventional induction motors.

When a motor is operated at a frequency higher than the design frequency, all reactances are increased by the ratio of frequencies. Inspection of Eq. (7.52) indicates that this causes the slip for which maximum torque is developed to become smaller; and if the frequency ratio is sufficiently high, the peak may be moved into the positive-speed region. Also, an increase in magnetizing reactance increases the dependence of $Z(S)$ on (S) . Therefore, a motor that does not single-phase at its design frequency may very well single-phase if operated at substantially higher frequencies. It is therefore not good practice to operate a servomotor at frequencies higher than the design frequency, even though the windings would not be damaged.

7.10. The Induction Motor as a Tachometer. When the reference phase of a two-phase induction motor is excited and when the rotor is turned by some outside source, an a-c voltage appears at the control-field terminals. This voltage may be used as an indication of speed, and a motor so used becomes a tachometer. When a motor is designed specifically for use as a tachometer, its rotor is of light construction in order to hold the added inertia to a minimum. For this reason, tachometers are usually built with drag-cup rotors. These have the additional advantage of having no slot ripple in the output, such as is observed with squirrel-cage rotors. The large air gap necessitated by the drag-cup construction is also an advantage, as will be seen presently.

The voltage generated may be obtained from the theory developed in the last section.¹ If the control winding is open-circuited, $Z_c = \infty$. As has already been demonstrated,

$$V_{M1} = V_M \frac{Z(S)}{Z(S) + Z(2 - S)} \quad (7.54)$$

$$V_{M2} = V_M \frac{Z(2 - S)}{Z(S) + Z(2 - S)} \quad (7.55)$$

The generated voltage is V_c and is given by

$$V_c = V_{c1} + V_{c2} = -jV_{M1} + jV_{M2} \quad (7.14)$$

$$\text{or} \quad V_c = -jV_M \frac{Z(S) - Z(2 - S)}{Z(S) + Z(2 - S)} \quad (7.58)$$

Equation (7.58) can be evaluated by making use of the expression for

¹ For an extensive treatment of the a-c tachometer, see R. H. Frazier, Analysis of the AC Tachometer by Means of Two-phase Symmetrical Components, *Trans. AIEE*, vol. 70, pp. 1894-1906, 1951.

$Z(S)$ given in Eq. (7.41). The result, after some reduction, becomes

$$V_c = -jV_M R'_2 Z_M^2 \frac{1 - S}{K_1 + K_2 S(2 - S)} \quad (7.59)$$

where

$$\begin{aligned} K_1 &\triangleq R'_2[(Z_1 + Z_M)(Z_M + R'_2 + jX'_2) + Z_1(Z_M + jX'_2)] \\ K_2 &\triangleq [Z_1 Z_M^2 - Z_1 X'^2_2 - Z_M X'^2_2 + jZ_M X'_2(Z_M + 2Z_1)] \\ Z_1 &\triangleq R_1 + jX_1 \\ Z_M &\triangleq jR_M X_M / (R_M + jX_M) \end{aligned}$$

By Eq. (7.2) $1 - S = \Omega/\Omega_s$; hence V_c is seen to be proportional to speed provided K_2 is negligible.

Since Eq. (7.59), as it stands, is rather complicated, we make use of the fact that most tachometers use drag-cup rotors, for which the rotor reactance is negligible¹ and the exciting impedance Z_M is relatively small and almost purely reactive. We suppose, therefore, that

$$\begin{aligned} X'_2 &= 0 \\ Z_M &\approx jX_M \ll R'_2 \\ Z_1 + Z_M &\approx Z_1 = R_1 + jX_1 \end{aligned} \quad (7.60)$$

Then Eq. (7.59) becomes approximately

$$V_c \approx jV_M \frac{X_M^2(\Omega/\Omega_s)}{(R_1 + jX_1)R'_2\{1 + [(X_M/R'_2)(\Omega/\Omega_s)]^2\}} \quad (7.61)$$

Note that the nonlinearity is of the “saturating” type; i.e., the voltage at high speeds is less than the value that would be obtained if the tachometer were linear. Note also that the phase angle between V_c and V_M depends on the power-factor angle of the stator and becomes 90° as the power-factor angle approaches zero.

The explanation of the operation of an a-c tachometer given above is based essentially on the so-called *double-revolving-field theory* of the operation of single-phase induction machines. While the results obtained are correct, a somewhat clearer physical picture of the operation is

obtained by use of the *cross-field theory*. According to this theory, the currents flowing in the main stator winding set up an alternating flux ϕ_M in the air gap (see Fig. 7.14). When the rotor rotates, two types of voltage are induced in it by the main flux, a transformer voltage E_M and a speed voltage E_Ω . This effect is explained in some detail in Sec. 4.4, and it is shown there that the speed voltage E_Ω is in time phase with the

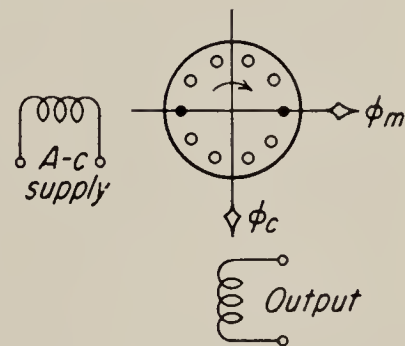


FIG. 7.14. Flux patterns in the a-c tachometer.

¹ Koopman, *op. cit.*, p. 322.

flux ϕ_M , is proportional to the speed of rotation, and is a maximum in the rotor conductors which are shown shaded in Fig. 7.14. The speed voltage results in a current flowing in the rotor, and this current, in turn, results in an alternating magnetic flux, called the *cross field*, ϕ_c . The cross-field flux then induces the output voltage in the turns of the output winding by ordinary transformer action. The speed voltage in the rotor, E_Ω , is directly proportional to speed if ϕ_M is constant. Thus the resulting current, the cross flux, and the output voltage should also be directly proportional to speed. The main flux ϕ_M may be assumed to be constant if the magnetic reaction of the rotor on the stator is negligible. This would correspond to a very small X_M and a large R_2 , and it is seen in Eqs. (7.59) and (7.61) that these are indeed the conditions for linearity.

Although a sudden change in speed results in an instantaneous increase of the speed voltage, it requires a short time for the resulting rotor current to be established unless the rotor reactance is negligible. This time delay results in a small time lag between the speed change and the resulting change of output voltage. It can be shown, however, that the resulting time constant is of the order of one period of the exciting frequency and is, therefore, normally negligible.

The primary advantage of the a-c tachometer over its d-c counterpart is the almost complete absence of commutator ripple. It is true that the output is an a-c voltage and must, therefore, usually be passed through a phase discriminator. However, for any speed, the output frequency is constant and thus can be filtered as much as is desirable. This is not possible with commutator ripple, since its frequency is proportional to speed, and a filter which works satisfactorily at a relatively high speed will be unsatisfactory at lower speeds. Additional advantages are the absence of brush friction and brush bounce. On the other hand, it is easier to keep a permanent-magnet d-c tachometer calibrated, particularly if the magnet has been properly aged. With a-c tachometers the output voltage is directly proportional to the reference voltage, and it therefore requires a closely regulated supply to maintain calibration. Furthermore, both R_1 and R_2 vary with changes of temperature, with immediate effect on calibration. Thus, when a tachometer is used to indicate absolute speed to a high accuracy, the d-c type is probably preferable, but for use in subsidiary stabilizing loops of servomechanisms, where calibration accuracy is relatively unimportant but where commutator ripple is highly objectionable, the a-c tachometer seems preferable.

7.11. Other A-C Servomotors. The Shaded-pole Induction Motor. Shaded-pole induction motors are used quite extensively to drive desk fans, phonograph turntables, and many other small-power devices in common use. Motors designed for these applications usually have a single, fixed shading coil per pole, often consisting simply of a heavy cop-

per ring. Neither their direction of rotation nor the developed torque can, therefore, be changed easily. To make shaded-pole motors reversible, they are built with two shading coils per pole. These coils are wound with a relatively large number of turns, and their ends are brought out to external terminals so that the current in the shading coils can be controlled. An exploded view of such a motor is shown in Fig. 7.15. Note that a standard squirrel-cage rotor is used.

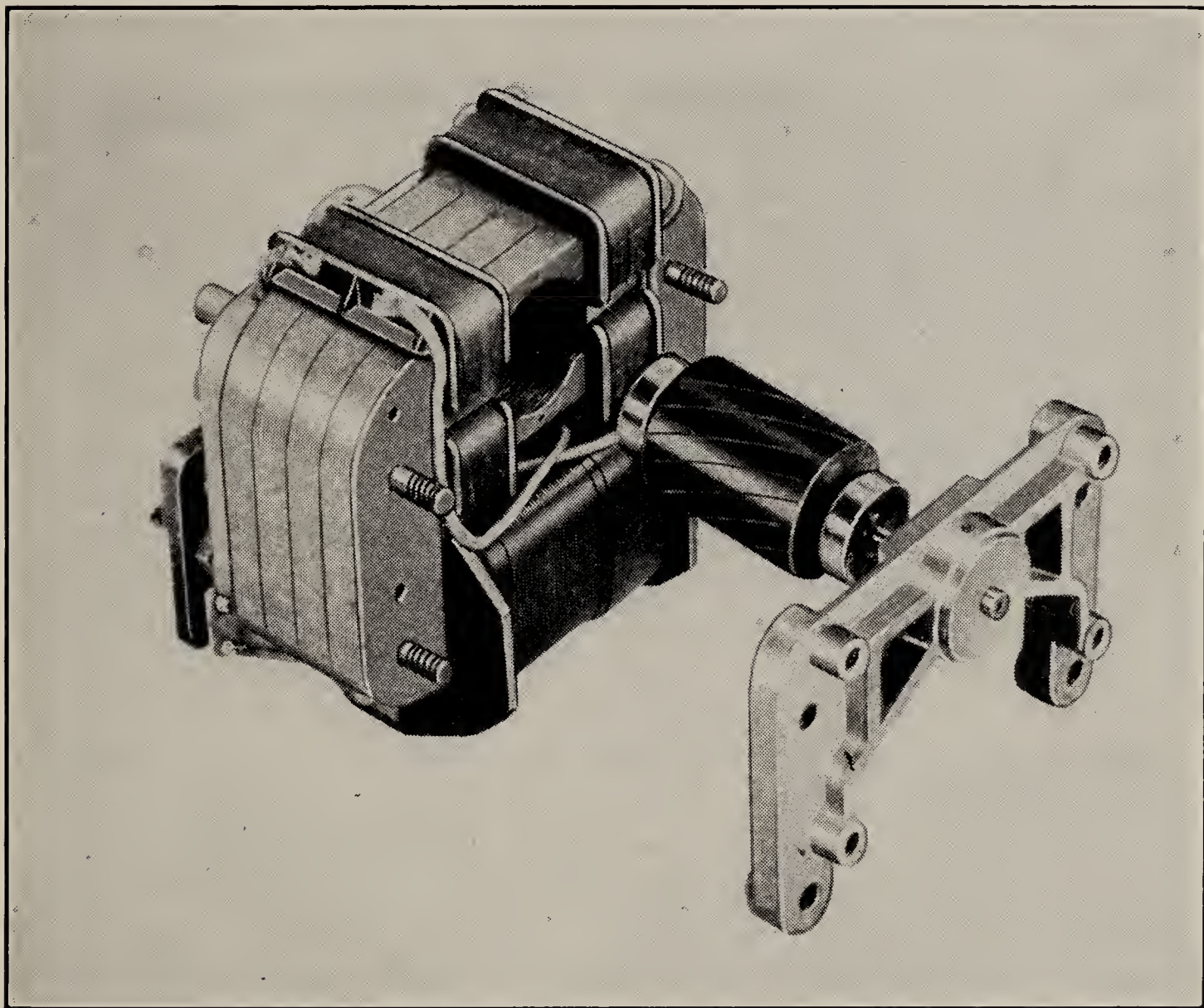


FIG. 7.15. Shaded-pole servomotor. (*Barber Colman Co.*)

The primary advantage of this motor is its very much lower cost compared to standard servomotors. This makes the motor useful in many noncritical commercial applications where low cost is a primary consideration. In such applications the motor is often used with relay control of the shading coils, and in this way it is possible to design simple follow-up systems employing no amplifiers and only the simplest and most rugged components. Somewhat higher performance can be obtained by electronic control of the shading-coil current. For best results the shading-coil current should be 90° out of phase with the current in the main field winding, and under these conditions the operation of the motor resembles that of a two-phase motor.

The performance of commercially available shaded-pole motors as measured by such criteria as the torque-to-inertia ratio or the torque-squared-to-inertia ratio is considerably poorer than that of conventional motors. Typically the torque-to-inertia ratio of a Kearfott type R111 motor is more than four times as high as that of a shaded-pole motor of comparable rating. Also, owing to the salient-pole construction of shaded-pole motors, the speed-torque curve tends to have irregularities not found in motors with well-distributed windings. Thus the shaded-pole motor is not ordinarily used in high-performance systems.

Qualitatively the operation of the motor is usually explained by noting

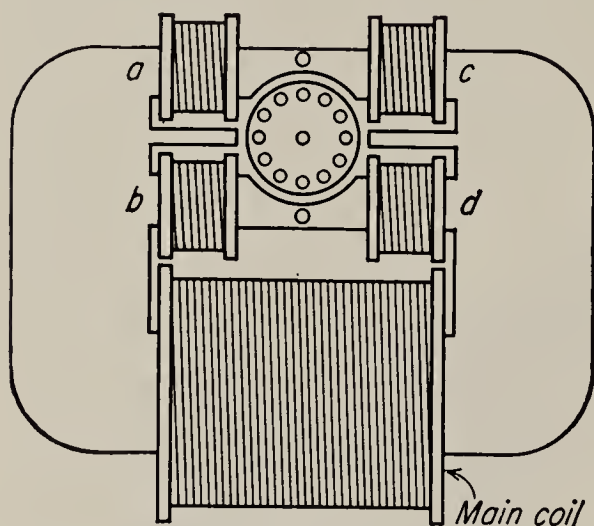


FIG. 7.16. Shaded-pole servomotor showing control coils.

that the action of the short-circuited shading coil delays the flux passing through it relative to the flux passing through the unshaded portion of the pole. This delay causes the peak of the flux wave to move from the unshaded to the shaded part of the pole, and this motion of the flux results in torque on the squirrel-cage winding as in the polyphase induction motor. Hence, if in Fig. 7.16 coils *a* and *d* are short-circuited, rotation will be in the clockwise direction; if *b* and *c* are short-circuited, the direction is reversed.

Quantitative analyses of the shaded-pole motor have been made by Trickey, Kron, and Chang.¹ Most of these analyses are rather complex, primarily because the salient-pole construction of shaded-pole motors makes an analysis based on sinusoidal air-gap flux distribution rather unrealistic. It is found empirically, however, that the speed-torque curves of many commercially available motors, when controlled by an electronic circuit, such as the one shown in Fig. 7.17, are quite regular and are similar to the characteristics of standard two-phase induction motors (see Fig. 7.18). Hence approximate transfer functions may be derived for shaded-pole motors in the same way as for two-phase motors (see Sec. 7.5). Since the applications in which these motors are used are not apt to be very critical, only the most approximate method seems justified. Thus, assume a straight-line relation for the speed-torque curve,

¹ P. H. Trickey, An Analysis of the Shaded-pole Motor, *Elec. Eng.*, September, 1936, pp. 1007–1014; Performance Calculation of Shaded-pole Motors, *Trans. AIEE*, vol. 66, pp. 1431–1438, 1947. G. Kron, Equivalent Circuits of the Shaded Pole Motor with Space Harmonics, *Trans. AIEE*, vol. 69, pp. 720–727, 1950. Chang, Equivalent Circuits and Their Applications in Designing Shaded-pole Motors, *Trans. AIEE*, vol. 70, pp. 690–698, 1951.

such as is shown by the dotted line in Fig. 7.18*a*; then by analogy with Eqs. (7.30), (7.38), and (7.39), we have

$$\hat{\Omega} = \frac{(\Omega_M/V_{c\max}) \hat{V}_c - (\Omega_M/Q_0) \hat{Q}}{J(\Omega_M/Q_0)s + 1}$$

(7.62)

where $V_{c\max}$ is the maximum voltage applied to the shading coil, J is the moment of inertia, V_c is the voltage applied to the shading coil, and Q is

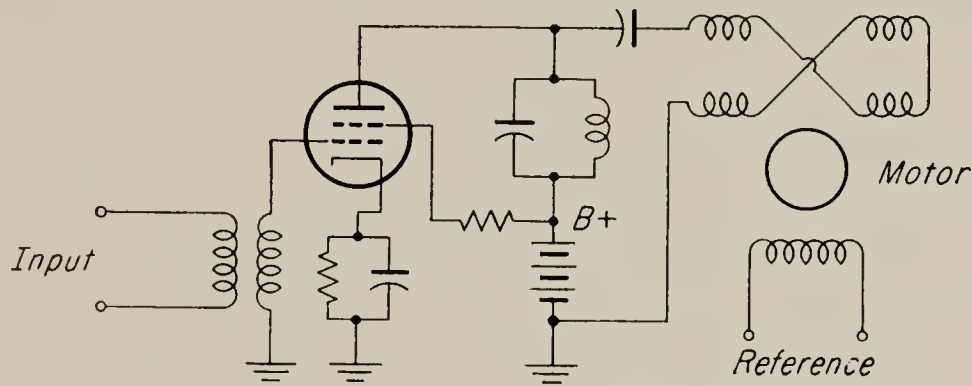


FIG. 7.17. Electronic control circuit used with shaded-pole motor.

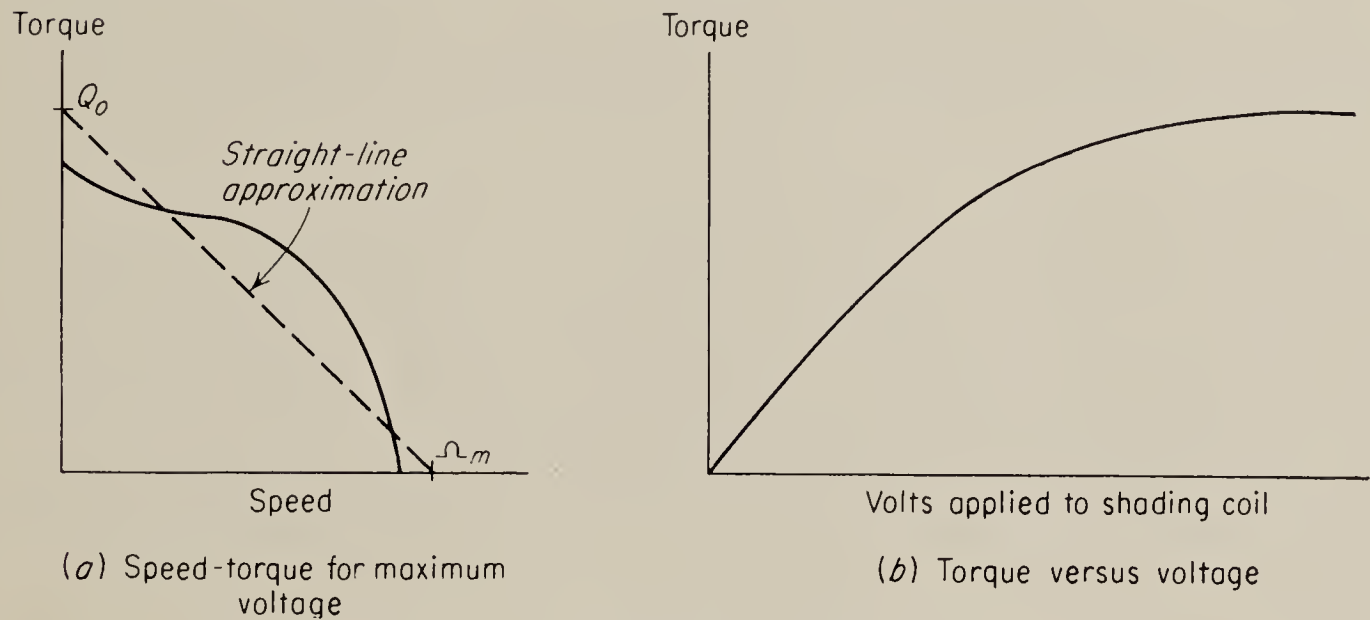


FIG. 7.18. Typical characteristics of shaded-pole motors.

the applied torque. Q_0 and Ω_M are the stalled torque and maximum speed, respectively, as determined by the straight-line approximation of Fig. 7.18*a*.

PROBLEMS

7.1. Plot the speed-torque curve for a two-phase induction motor having the equivalent circuit shown in Fig. 7.19. The voltage per phase is 50 volts, and the two voltages are 90° out of phase. The frequency is 60 cps, and the motor is wound for two poles.

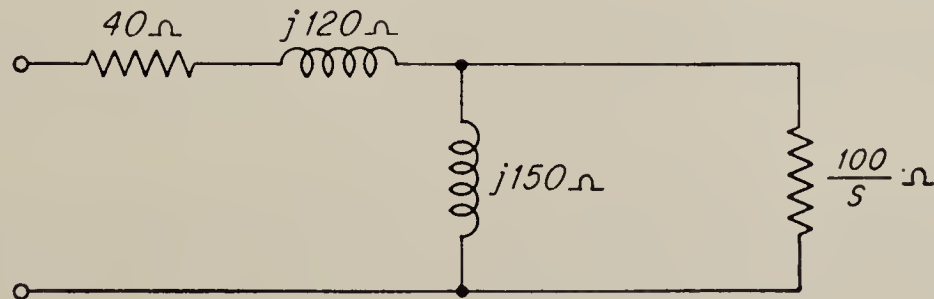


FIG. 7.19

7.2. Assume that one phase of the motor of Prob. 7.1 is supplied with a fixed 60-cps voltage of 50 volts and that the variable voltage applied to the other phase is 90° out of phase with the fixed voltage. Plot speed-torque curves for variable voltages of 0, 10, 20, 30, and 40 volts.

7.3. Find the constants C_1 and C_2 [see Eq. (7.28)] for the motor of Prob. 7.1.

7.4. Find the approximate motor transfer function for the motor of Prob. 7.1 on the assumption that the moment of inertia of the rotor is 0.1 oz-in.² Consider two operating points: (a) speed = zero; variable voltage = zero; (b) speed = zero, variable voltage = fixed voltage = 50 volts. Variable voltage lags fixed voltage by 90°.

7.5. The following data are available on the Diehl FPE-25-11 motor:

Output, watts.....	5
Poles.....	2
Reference volts.....	115
Control volts.....	115
Frequency, cps.....	60
Locked torque, oz-in.....	5.5
Torque at 1,800 rpm, oz-in.....	3.8
Moment of inertia, oz-in. ²	0.098

Assume that the torque equation for 115 volts on both phases is given by

$$T = \frac{S}{C_1 + C_2S^2} \qquad S = \text{slip}$$

(a) Find the quasi-linear transfer function for the operating point of zero speed, zero control voltage. (b) Find the quasi-linear transfer function for the operating point of zero speed, full control voltage. (c) Find the approximate transfer function using the straight-line approximation to the speed-torque curves. The control voltage is 90° out of phase with the reference voltage in all cases.

7.6. Assume that the motor of Prob. 7.1 is used as an a-c tachometer. If the reference voltage is 50 volts, find the numerical relation between velocity and output voltage.

CHAPTER 8

MECHANICAL NETWORKS AND GEARS

8.1. Introduction. Most feedback control systems have some mechanical components, as distinguished from electric and hydraulic components, etc. However in this discussion the classification *mechanical element* will be restricted to elements that actually perform a functional service in the loop such as sensing speed or performing addition or subtraction. The advantages of mechanical components are reliability, environmental stability, and, surprisingly, sometimes size. Certain mechanical filters have been constructed that are smaller than the equivalent electric filter and superior to the electric filters in performance. In general, however, mechanical elements are heavier and bulkier than other types of elements, although if the power supplies required for pneumatic, electric, and hydraulic systems are included in the weight and size calculations, mechanical systems appear to better advantage. We shall consider in this chapter simple arrangements of mechanical elements and in the following chapter somewhat more elaborate configurations.

8.2. Springs, Masses, and Dashpots. The equation that relates force and mass may be written as

$$\mathcal{F} = Ma = M \frac{d^2x}{dt^2} \quad (8.1)$$

where a is acceleration in the x direction. The Laplace transform of this relation is

$$\hat{\mathcal{F}} = Ms^2\hat{x} \quad (8.2)$$

The simple spring is defined as a device in which the force is linearly related to the deflection:

$$\hat{\mathcal{F}} = K\hat{x} \quad (8.3)$$

In the simple dashpot, the reaction force is proportional to relative velocity between its terminals:

$$\hat{\mathcal{F}} = Bs\hat{x} \quad (8.4)$$

Figure 8.1 gives the conventional representation for these devices.

While mechanical systems may be analyzed directly, it is occasionally useful to employ a mechanical-to-electrical analogue for ease in solution

of mechanical systems. Analyzing mechanical systems in terms of electrical quantities permits use of the extensive work done on electric filters and is convenient if a computer is available to implement the solution.

In Fig. 8.2 are shown three systems, two electric and one mechanical, with their describing equations. Since all the equations are of the same form, their solutions are identical.

The mechanical system may be compared with either of the electric networks. The series or impedance analogue compares the mechanical

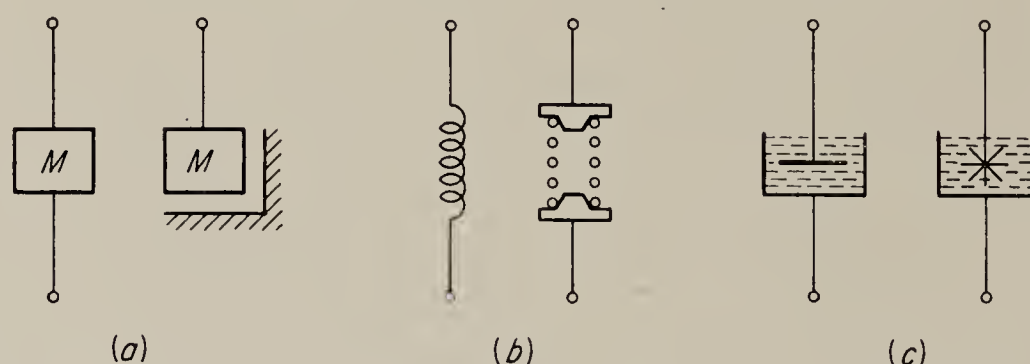


FIG. 8.1. Conventional representation of mass, spring, and dashpot.

system with the first of the electric networks. Mechanical force is then analogous to voltage. The second, and more modern, analogue is the mobility analogue, in which mechanical force is analogous to current. As Firestone¹ points out, the mobility analogue is the more convenient of the two because the schematic diagrams look the same. Furthermore, current is more closely allied to force than is voltage, since the methods of measurement are similar. In Table 8.1 is shown a comparison of electrical quantities and analogous mechanical quantities.

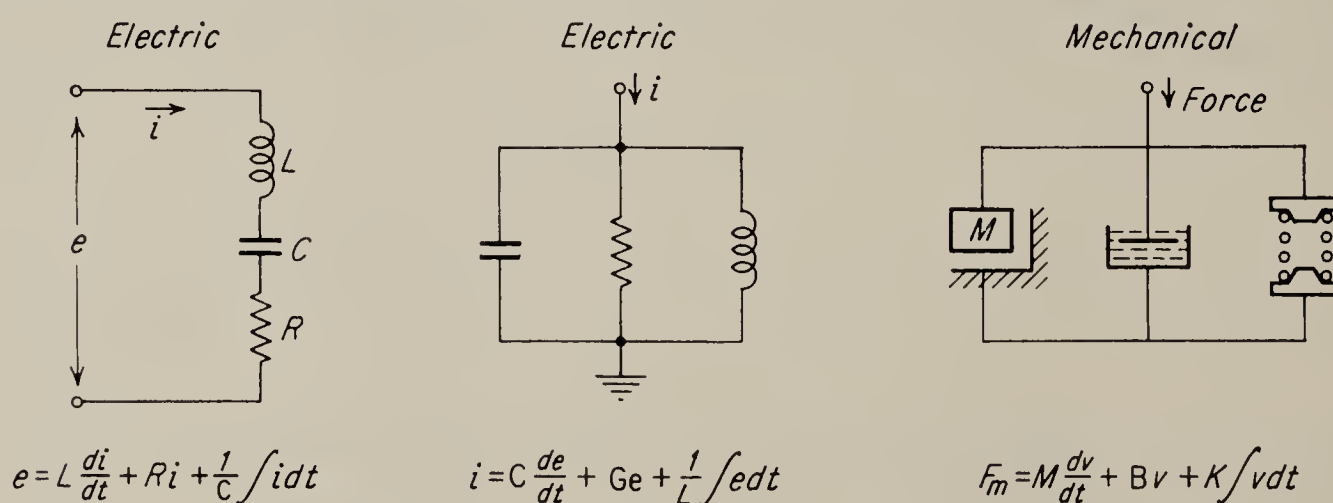


FIG. 8.2. Three analogous circuits with their equations.

By analogy to electric circuits or directly from the law of conservation of energy, we may write the relations for mechanical elements. Mechanical elements connected in parallel must have the same velocity and position across them, and elements in series must have the same force through

¹ F. A. Firestone, The Mobility Method of Computing the Vibration of Linear Mechanical and Acoustical Systems: Mechanical-Electrical Analogies, *J. App. Phys.*, vol. 9, p. 373, 1938.

TABLE 8.1. A COMPARISON OF ELECTRICAL AND MECHANICAL QUANTITIES

Mechanical	Electrical (mobility analogue)	Electrical (series analogue)
Force, \mathcal{F}	Current, i	Voltage, e
Velocity, v	Voltage, e	Current, i
Mass, M	Capacitance, C	Inductance, L
Dashpot (viscous damping), B	Resistance, $G = \frac{1}{R}$	Resistance, R
Spring, $S = 1/K$	Inductance, L	Capacitance, C

them. We may add the velocities of components in series in order to find the total velocity. As an example, consider the transfer function from the input to output of the elements shown in Fig. 8.3. Since the elements are in series the same force must be transmitted through each, if they are to be in equilibrium.

Thus:

$$\hat{\mathcal{F}}_{\text{total}} = \hat{\mathcal{F}}_{\text{dashpot}} = \hat{\mathcal{F}}_{\text{spring}}$$

or

$$\hat{\mathcal{F}}_{\text{total}} = Bs(\hat{x} - \hat{y}) = K\hat{y} \tag{8.5}$$

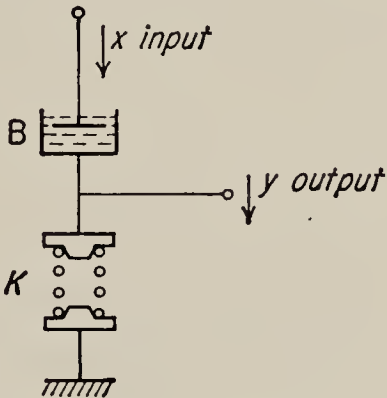


FIG. 8.3. Dashpot and spring in series.

Since the two elements are in series, their relative velocities may be added to obtain the total velocity of the point x . By relative velocities is meant the velocity across each element, or the velocity that one end of the element has compared with its own other end.

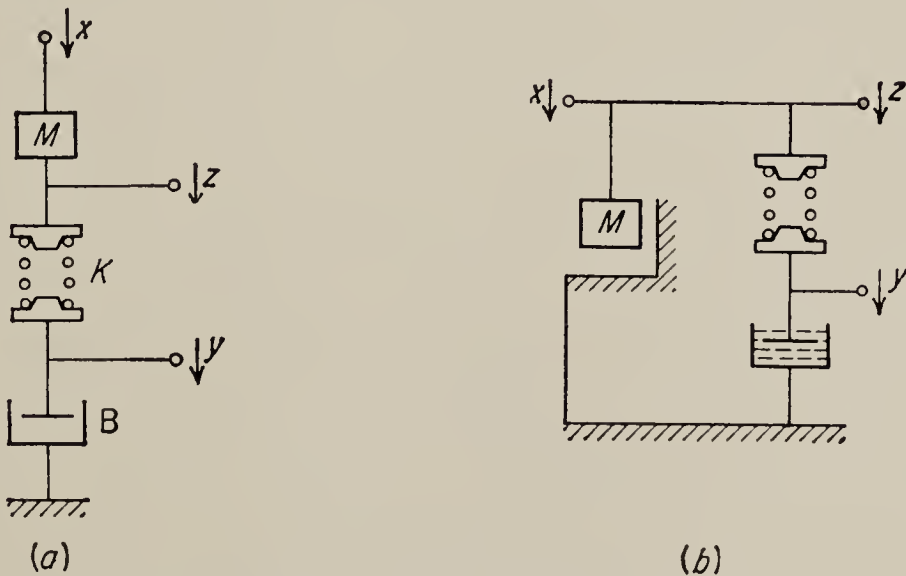


FIG. 8.4. Series mass, spring, and dashpot.

The transfer function of the system shown in Fig. 8.4 is most easily obtained by rearranging Eq. (8.5) as follows:

$$Bs\hat{x} = (Bs + K)\hat{y} \tag{8.6}$$

so that

$$\frac{\hat{y}}{\hat{x}} = \frac{Bs}{Bs + K} = \frac{(B/K)s}{(B/K)s + 1} \quad (8.7)$$

One more example may clarify a point concerning the handling of mass. In Fig. 8.4a is shown a series-connected mass, spring, and dashpot. It should be apparent that there can be no relative velocity across the mass. One end of the mass is moving at exactly the same velocity as the other end. The mass moves only relative to ground; this is the implication of the schematic representation of mass shown in Fig. 8.4b, which is identical to Fig. 8.4a. Thus the mass in no way affects the transfer function of the system for position or velocity.¹ This may be more clearly seen by redrawing the circuit as shown in Fig. 8.4b. We may find the transfer function in exactly the same manner as above.

$$\frac{\hat{y}}{\hat{x}} = \frac{1}{(B/K)s + 1} \quad (8.8)$$

If the relationship between input force and output force is desired, a load must be assumed and the mass must then be considered.

8.3. Mechanical Equalizers. In addition to their use as the usual system elements, springs and dashpots are also sometimes installed in systems in order to modify the system transfer function to improve closed-loop operation. These combinations are called *controllers*, or *equalizers*, and are analogous to their electrical counterparts.

Figure 8.5 is a table of transfer functions of such equalizers using only springs and dashpots.² Note that several of the configurations result in the same transfer function. This adds flexibility to the design, since it may be possible to utilize a portion of the original system in the control network and several choices of configuration increase this possibility.

There is no theoretical reason for not extending the use of mechanical elements to the more complicated networks such as the bridged T and twin T used in carrier servos. Figure 8.6 shows a bridged-T network. Figure 8.6a is the electric network, and Fig. 8.6b is its mechanical analogue. The twin T can likewise be constructed with mechanical components. Figure 8.7 shows this configuration.³ Since complete design data are available for the electric networks (see Chap. 1) used in carrier systems,

¹ Electrical engineers who wish to compare this circuit to an electrical analogue should remember that since the mass can show relative velocity only with respect to ground, it is equivalent to a capacitor connected to ground. The electrical analogue of the circuit is a capacitor in parallel with a coil and resistor in series.

² J. E. Gibson, *Fourteen Ways to Construct Control Functions Mechanically*, *Control Eng.*, vol. 2, pp. 65–69, May, 1955.

³ Truxal, "Automatic Feedback Control System Synthesis," McGraw-Hill Book Company, Inc., New York, 1955, p. 396.

Schematic diagram	Transfer function*	α diagram, log gain magnitude vs. log frequency in radians/sec (asymptotic)
<div>(1)<div></div></div>	<div>$\frac{\hat{Y}}{\hat{X}} = \frac{Ts}{1 + Ts}$$T \triangleq \frac{B}{K}$</div>	<div></div>
<div>(2)<div></div></div>	<div>$\frac{\hat{Y}}{\hat{X}} = \frac{1}{1 + Ts}$$T \triangleq \frac{B}{K}$</div>	<div></div>
<div>(3)<div></div></div>	<div>$\frac{\hat{Y}}{\hat{X}} = \frac{T_2}{T_1} \frac{1 + T_1 s}{1 + T_2 s}$$T_1 \triangleq \frac{B_1}{K_1} \quad T_2 \triangleq \frac{B_1}{K_1 + K_2}$</div>	<div></div>
<div>(4)<div></div></div>	<div>$\frac{\hat{Y}}{\hat{X}} = \frac{K_1}{K_1 + K_2} \frac{1}{1 + Ts}$$T \triangleq \frac{B_2}{K_1 + K_2}$</div>	<div></div>
<div>(5)<div></div></div>	<div>$\frac{\hat{Y}}{\hat{X}} = \frac{1 + T_2 s}{1 + T_1 s}$$T_1 \triangleq \frac{B_1 + B_2}{K_1} \quad T_2 \triangleq \frac{B_1}{K_1}$</div>	<div></div>
<div>(6)<div></div></div>	<div>$\frac{\hat{Y}}{\hat{X}} = \frac{T_1 s}{1 + T_2 s}$$T_1 \triangleq \frac{B_1}{K_2} \quad T_2 \triangleq \frac{B_1 + B_2}{K_2}$</div>	<div></div>

FIG. 8.5. Table of mechanical equalizers.

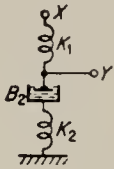
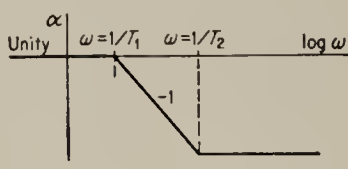
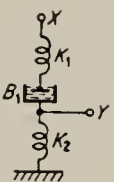
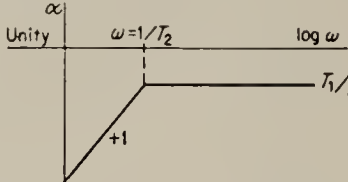
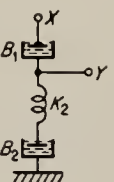
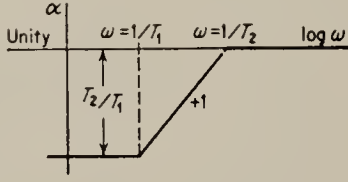
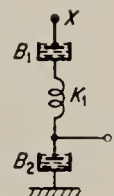
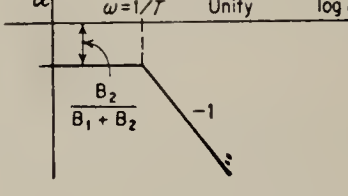
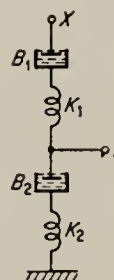
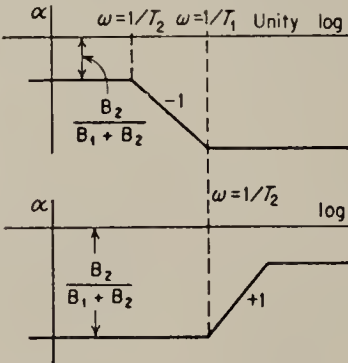
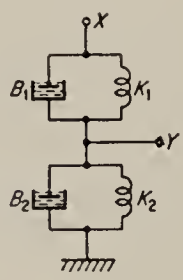
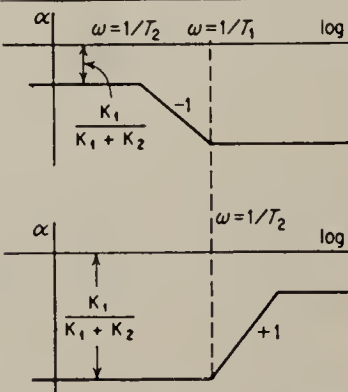
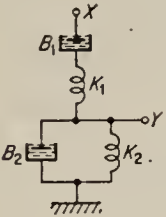
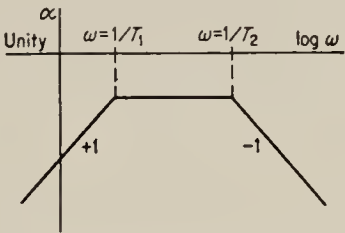
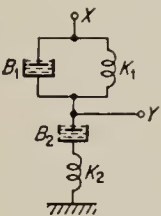
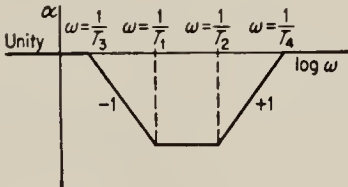
Schematic diagram	Transfer function*	α diagram, log gain magnitude vs. log frequency in radians/sec (asymptotic)
<div>(7) </div>	$\frac{\hat{Y}}{\hat{X}} = \frac{1 + T_2 s}{1 + T_1 s}$ $T_1 \triangleq \frac{B_2}{K_1} + \frac{B_2}{K_2} \quad T_2 \triangleq \frac{B_2}{K_2}$	
<div>(8) </div>	$\frac{\hat{Y}}{\hat{X}} = \frac{T_1 s}{1 + T_2 s}$ $T_1 \triangleq \frac{B_1}{K_2} \quad T_2 \triangleq \frac{B_1}{K_1} + \frac{B_1}{K_2}$	
<div>(9) </div>	$\frac{\hat{Y}}{\hat{X}} = \frac{T_2}{T_1} \frac{1 + T_1 s}{1 + T_2 s}$ $T_1 \triangleq \frac{B_2}{K_2} \quad T_2 \triangleq \frac{B_1 B_2}{K_2 (B_1 + B_2)}$	
<div>(10) </div>	$\frac{\hat{Y}}{\hat{X}} = \frac{B_1}{B_1 + B_2} \frac{1}{1 + T s}$ $T \triangleq \frac{B_1 B_2}{K_1 (B_1 + B_2)}$	
<div>(11) </div>	$\frac{\hat{Y}}{\hat{X}} = \frac{B_1}{B_1 + B_2} \frac{1 + T_1 s}{1 + T_2 s}$ $T_1 \triangleq \frac{B_2}{K_2} \quad T_2 \triangleq \frac{(K_1 + K_2) B_1 B_2}{K_1 K_2 (B_1 + B_2)}$ <p>See Note 1</p>	
<div>(12) </div>	$\frac{\hat{Y}}{\hat{X}} = \frac{K_1}{K_1 + K_2} \frac{1 + T_1 s}{1 + T_2 s}$ $T_1 \triangleq \frac{B_1}{K_1} \quad T_2 \triangleq \frac{B_1 + B_2}{K_1 + K_2}$ <p>See Note 1</p>	

FIG. 8.5 (Continued)

Schematic diagram	Transfer function*	α diagram, log gain magnitude vs. log frequency in radians/sec (asymptotic)
<div>(13)</div> <div></div>	<div>Approximate (see Note 2):</div> $\frac{\hat{Y}}{\hat{X}} = \frac{T_3 s}{(1 + T_1 s)(1 + T_2 s)}$ <div>Exact:</div> $\frac{\hat{Y}}{\hat{X}} = \frac{T_3 s}{1 + (T_1 + T_4)s + T_1 T_2 s^2}$ $T_1 \triangleq B_1 \left(\frac{1}{K_1} + \frac{1}{K_2} \right)$ $T_2 \triangleq \frac{B_2}{K_1 + K_2}$ $T_3 \triangleq \frac{B_1}{K_2} \quad T_4 \triangleq \frac{B_2}{K_2}$	<div></div> <div>When $T_1 > T_2$ breaks are as shown. If $T_2 > T_1$ reverse labels on breaks.</div>
<div>(14)</div> <div></div>	<div>Approximate (see Note 3):</div> $\frac{\hat{Y}}{\hat{X}} = \frac{(1 + T_1 s)(1 + T_2 s)}{(1 + T_3 s)(1 + T_4 s)}$ $T_1 \triangleq \frac{B_1}{K_1} \quad T_2 \triangleq \frac{B_2}{K_2}$ $T_3 \triangleq \frac{B_1 + B_2}{K_1} \quad T_4 \triangleq \frac{B_1 B_2}{(B_1 + B_2) K_2}$	<div></div>

* In the steady state $s = j\omega$.

NOTE 1. By choice of parameters T_2 can be made larger or smaller than T_1 , and thus either a -1 slope or a $+1$ slope can be synthesized.

NOTE 2. The approximation is good for $T_1 \gg T_2$. To find the approximate relation B_2 is considered open at low frequencies, and B_1 is considered shorted at high frequencies. See Chap. 1 for discussion of approximate design methods.

NOTE 3. The approximate relation is obtained by considering K_2 shorted at low frequencies and K_1 open at high frequencies. Multiplied out, the approximate relation is

$$\frac{\hat{Y}}{\hat{X}} = \frac{\left(1 + \frac{B_1}{K_1} s\right) \left(1 + \frac{B_2}{K_2} s\right)}{1 + \left(\frac{B_1}{K_1} + \frac{B_2}{K_1} + \frac{B_1 B_2}{(B_1 + B_2) K_2}\right) s + \frac{B_1 B_2}{K_1 K_2} s^2}$$

while the exact relation is

$$\frac{\hat{Y}}{\hat{X}} = \frac{\left(1 + \frac{B_1}{K_1} s\right) \left(1 + \frac{B_2}{K_2} s\right)}{1 + \left(\frac{B_1}{K_1} + \frac{B_2}{K_1} + \frac{B_2}{K_2}\right) s + \frac{B_1 B_2}{K_1 K_2} s^2}$$

The approximation is therefore good as long as $T_3 \gg T_2$.

FIG. 8.5 (Continued)

it is probably simplest to design the mechanical networks by analogy to the electric networks.

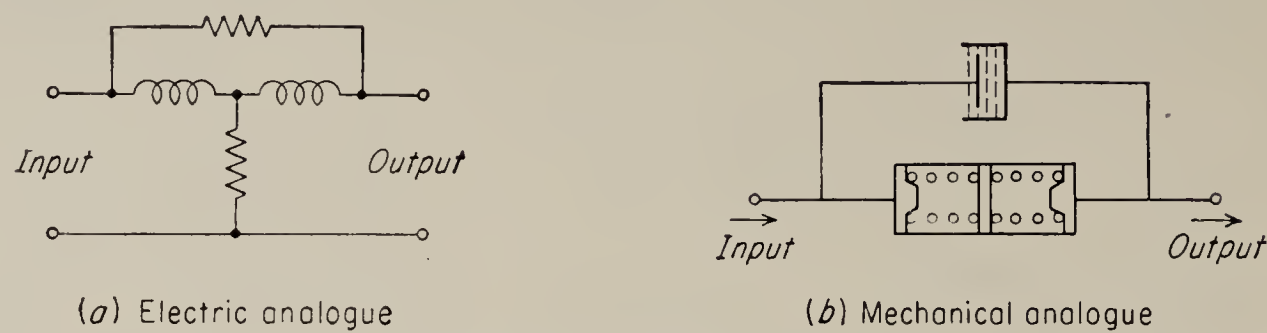
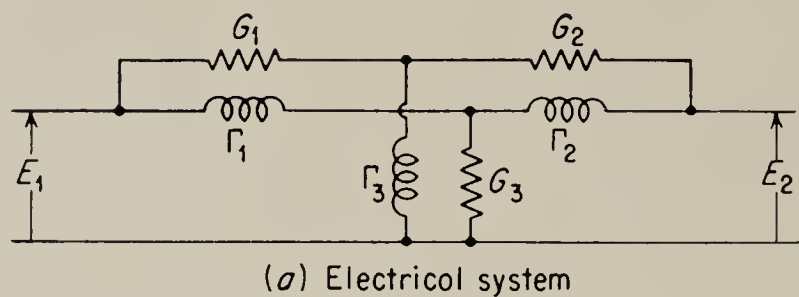
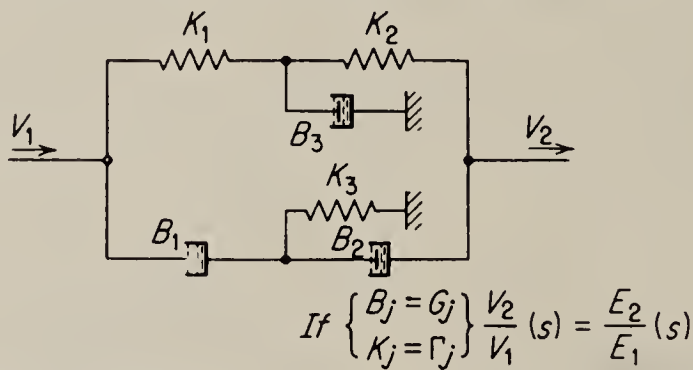


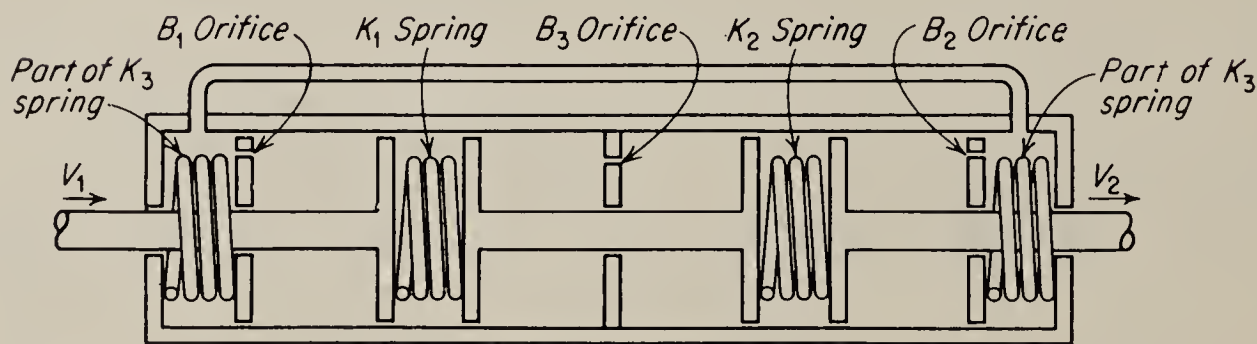
FIG. 8.6. Bridged T.



(a) Electrical system
($\Gamma = \frac{1}{L}$, $G = \frac{1}{R}$; Γ and G in inverse henrys and mhos)



(b) Analogous mechanical system (translational motion only)



(c) Possible mechanical unit for circuit of (b)
(system filled with hydraulic fluid)

FIG. 8.7. Mechanical twin-T network. (*Truxal*)

8.4. Electromechanical Networks for Carrier Servos. A-c servomechanisms or carrier-type systems are simple to manufacture and usually require smaller, lighter components than do d-c systems. A-c systems, however, usually cannot match the high performance of d-c systems because of the difficulty of equalizing a-c systems. We have discussed lead networks for a-c systems in Chap. 1. It will be recalled that these networks do not produce as much phase lead as a conventional lead equalizer in an equivalent d-c system. In addition to the smaller amount

of lead available, the conventional carrier networks are quite sensitive to shifts in the carrier frequency. A shift in carrier frequency will reduce the theoretical phase lead and may, in fact, produce phase lag. This sensitivity to carrier shift is on an absolute basis rather than on a percentage basis. Thus in aircraft installations, where the simplicity of an a-c servo is most appreciated, a given percentage shift in frequency of the

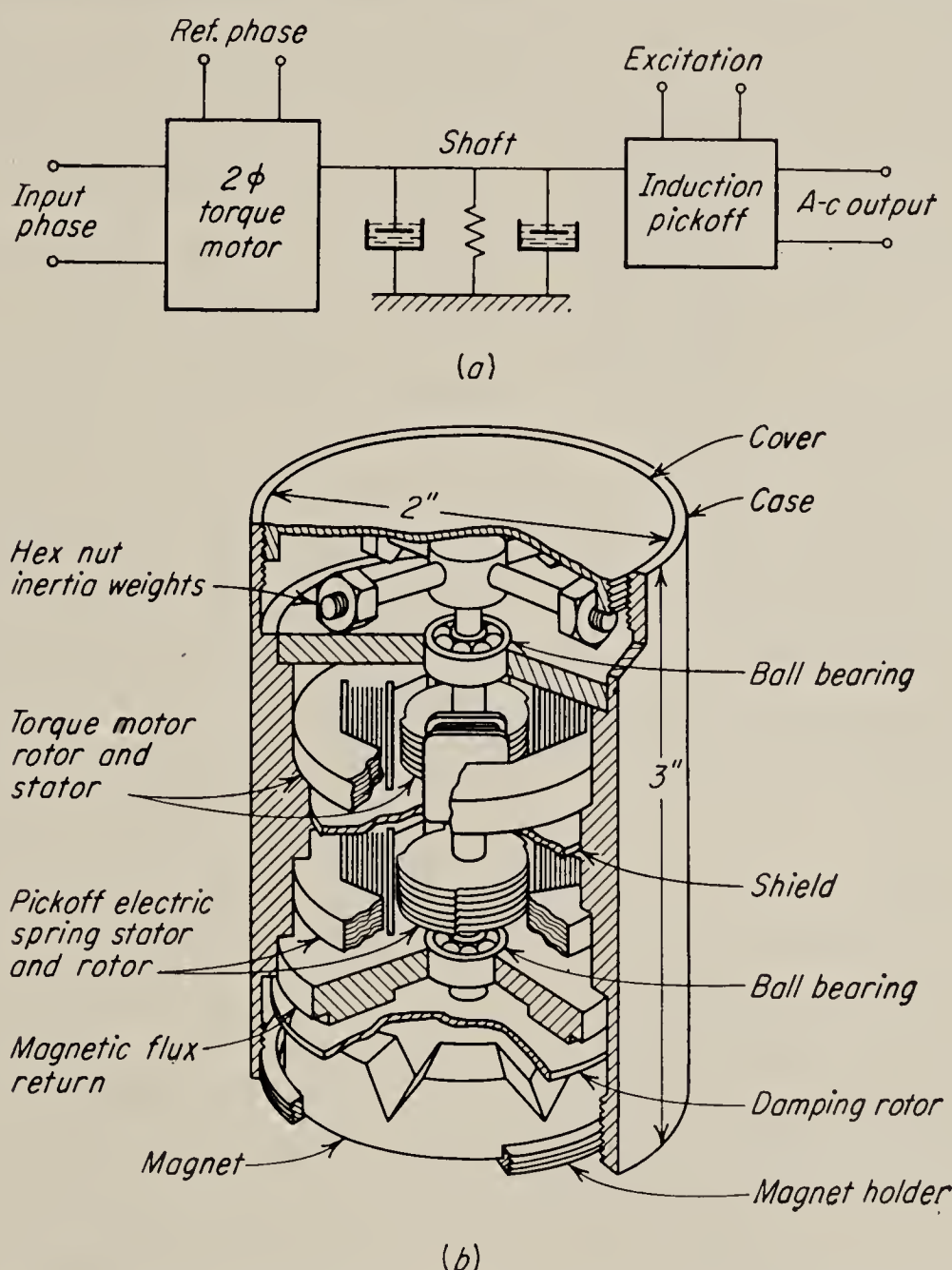


FIG. 8.8. Electromechanical lead network. (From McDonald)

high-frequency carrier (400 to 800 cps) is more damaging than the same percentage shift would be in a 60-cycle carrier system.

McDonald¹ suggests that a compact electromechanical unit may be used to demodulate the a-c signal and apply it to a mechanical d-c lead network and then convert the signal back into a-c form. The block diagram of such a device is shown in Fig. 8.8a, and in Fig. 8.8b is shown a cutaway drawing of the actual device. The torque motor acts

¹ D. McDonald, Electromechanical Lead Networks for A-C Servomechanisms, *Rev. Sci. Instr.*, vol. 20, p. 775, 1949.

as the demodulator, and the induction pick-off acts as the modulator. It will be noted that such a unit is light, compact, and completely free from effects of carrier-frequency shift. Furthermore the full phase lead available from d-c networks may be obtained. It should be noted that this is simply an electromechanical form of the demodulator-d-c-equalizer-

modulator chain discussed in Chap. 6.

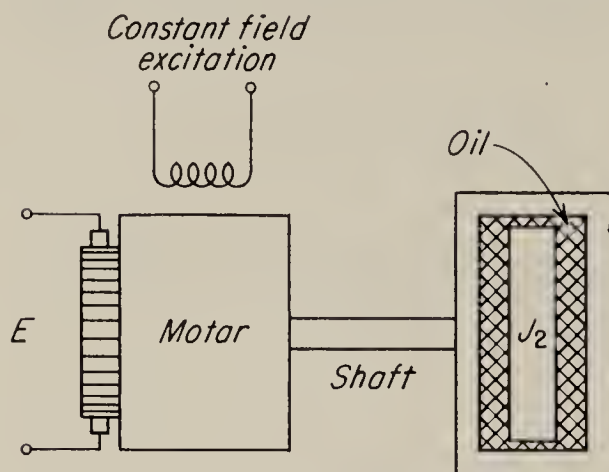


FIG. 8.9. The Lancaster damper.

8.5. The Lancaster Damper. A rather common form of mechanical equalizer is known as the *Lancaster damper*. It consists of a damper plate mounted on the shaft of a motor and a second plate which is free to turn with respect to the first plate, as shown in Fig. 8.9. The intervening space is filled with damping fluid, usually hydraulic oil.

A second type of damper, shown in Fig. 8.10, works on an electromagnetic principle. The copper cup fixed to the motor shaft is quite light and adds a negligible amount to the original inertia of the motor. The toothed steel wheel is magnetized. The flux passes from a north pole across the air gap to the steel ring and back across the air gap to the adjacent south pole. When the copper cup moves in the air gap with

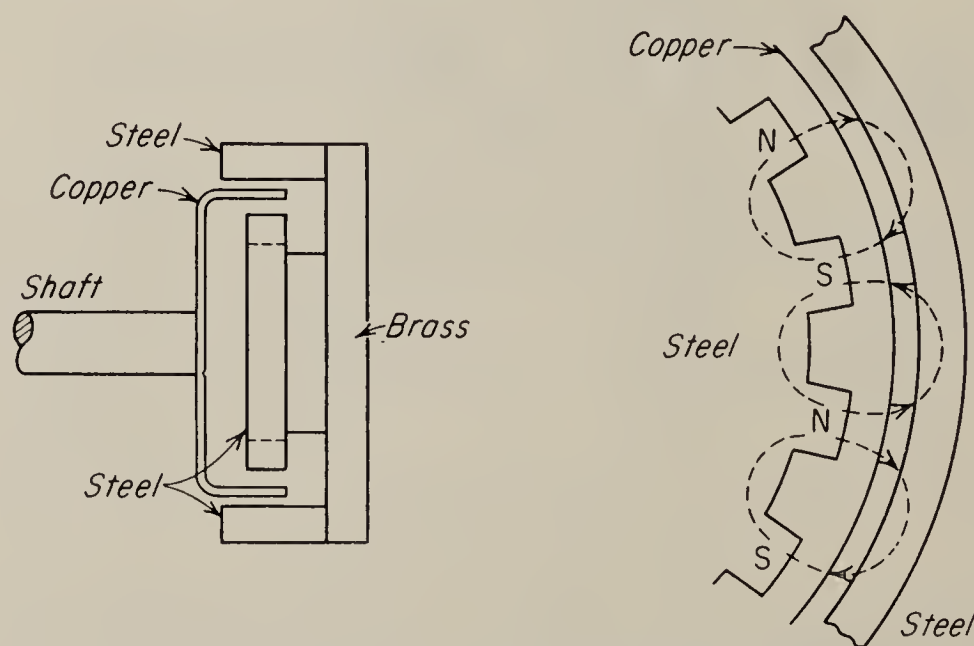


FIG. 8.10. A magnetic form of Lancaster damper.

respect to the flux, eddy currents are set up in the copper, which in turn produce a retarding torque.

The frequency response of a d-c motor with an inertia load and with no damper has been discussed in Chap. 4 and its asymptotic α diagram is shown in Fig. 8.11a. The break occurs at a frequency of $k_v k_t / J_1 R_a$ radians/sec, where k_v is the generated voltage constant of the motor, k_t is

the torque constant of the motor, J_1 is the moment of inertia of the rotor and load, and R_a is the armature resistance. The reader may show that, with the coefficients of the damper properly selected, the new α diagram will be as shown in Fig. 8.11b. The response is thus well damped. The empirical design of these dampers long preceded the control-system approach to the problem and has been used with equal success with a-c motors. The damper is used in many simple a-c or d-c systems as the only equalizer in the loop. In certain cases the use of this device makes possible a satisfactory response in a very simple a-c control system, where without the damper a more complex d-c system with an electronic equalizer would be required in order to meet specifications.

The Lancaster damper has several disadvantages. Since it is at a high-power-level point in the loop, it loads the motor and may absorb an appreciable portion of its rated torque. Another disadvantage is that, after the mass J_2 (see Fig. 8.9) has reached an appreciable speed, it provides a torque that tends to cause the motor to overshoot the zero-error position; this may cause poor synchronizing performance. Also, in the hydraulic-type damper the effect of temperature on the coefficient of viscous damping will change the frequency of the breaks on the α diagram and thus can decrease the effective damping.

A modification of the Lancaster, or *untuned*, damper is the so-called *tuned damper*.¹ When, in addition to the viscous-fluid coupling, the free plate is coupled to the shaft by a spring, the damper is said to be tuned.

8.6. Gears. Gears are used to reverse the direction of rotation of a shaft, to provide self-locking action, to provide a right-angle drive, etc. The most common use is to provide a change in shaft speed. For instance, most electric motors are designed to operate at relatively high speeds and low torque. The typical control application for motors is just the reverse of this. A reduction in shaft speed and an increase in shaft torque are the usual application for gearing between an electric motor and its shaft load.

The *spur gear and pinion* is the most common type of gearing in control

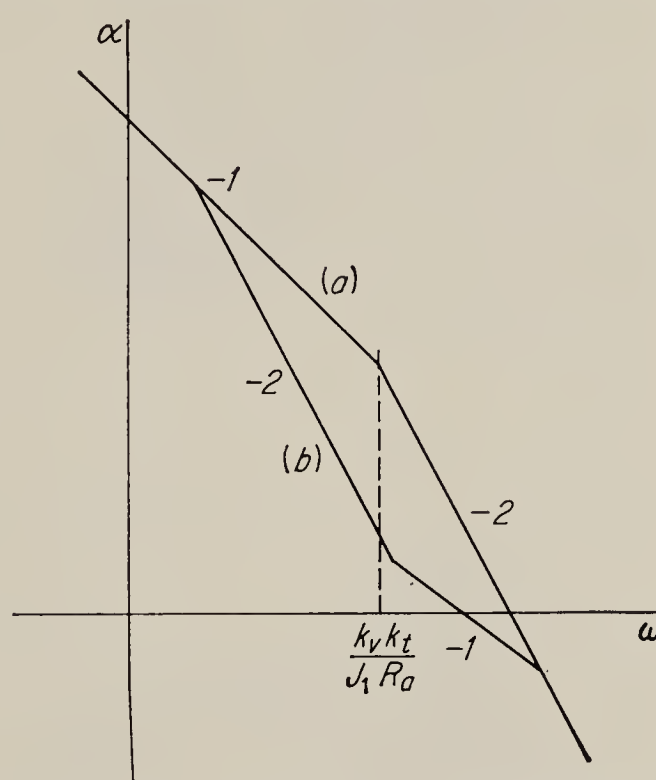


FIG. 8.11. Asymptotic α diagram of motor with and without damper.

¹ Greenwood, Holdam, and MacRae, "Electronic Instruments," Radiation Laboratory Series, vol. 21, McGraw-Hill Book Company, Inc., New York, 1948, Sec. 11.6.

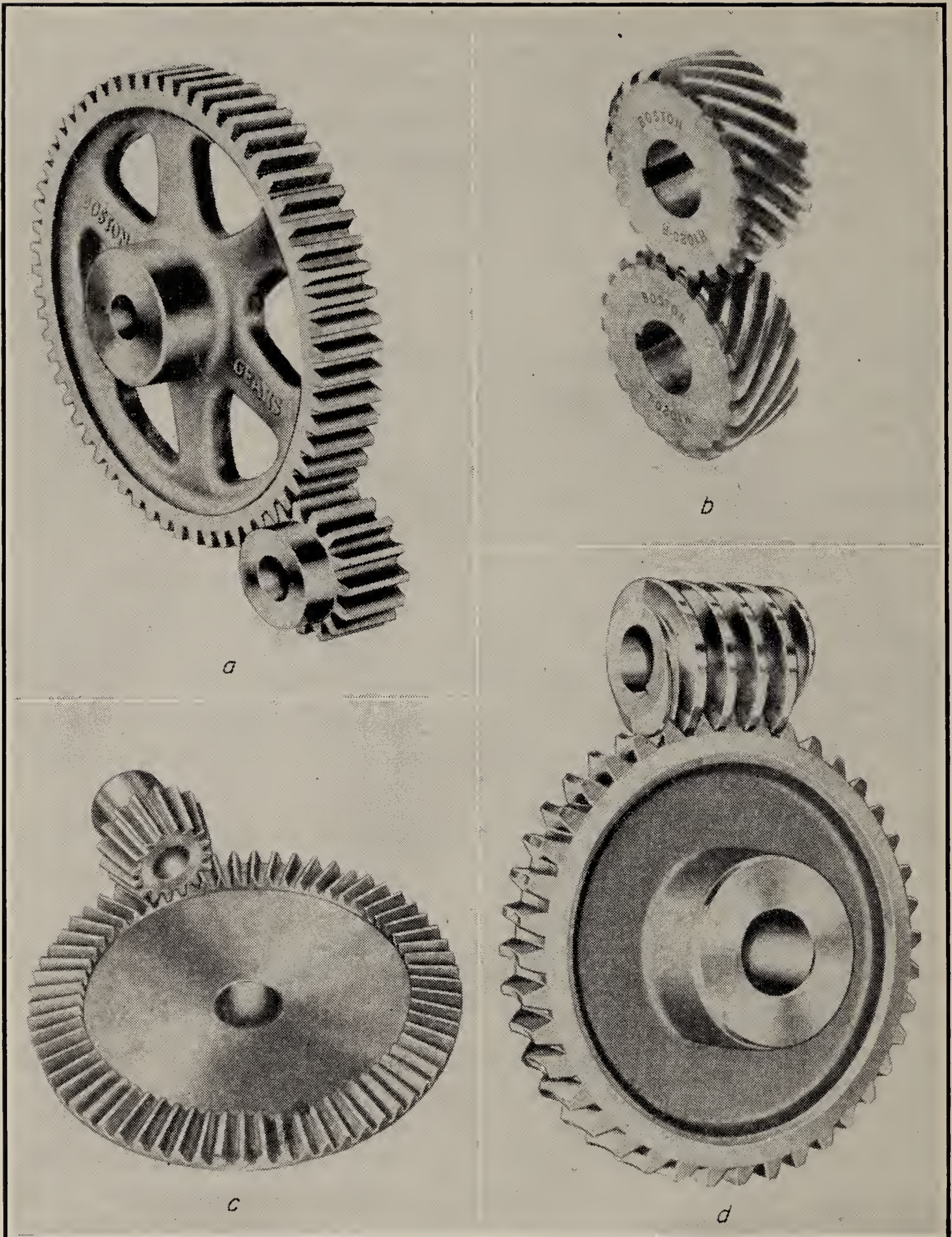


FIG. 8.12. (a) Spur gear and pinion; (b) helical gear; (c) bevel gear and pinion; (d) worm and gear.

applications.¹ The small gear is called the pinion (see Fig. 8.12). This type of gearing is simple to manufacture, since the gear teeth are at right

¹ For more complete information on the design of gears see Berard, Waters, and Phelps, "Principles of Machine Design," The Ronald Press Company, New York, 1955.

angles with the body, or gear disk. The spur gear is a high-efficiency gear and is thus reversible. Generally the highest gear ratio that is practical with spur gears is about 10:1 for one mesh. Figure 8.13 gives a comparison of proportions for the American Standard 20°-involute fine-pitch system and the 14½°-pressure-angle system as modified for fine-pitch service.¹

For many years gear design has been standardized on these two pressure angles, 14½° and 20°. The 14½° design is the older of the two standards and was the product of practical considerations rather than engineering research.² As modified for fine-pitch service, the 14½° design has an undercut pinion, which decreases the gear strength. The smaller pressure-angle design theoretically provides smaller friction and less backlash, but in practice the improvement is negligible. In addition, tests show

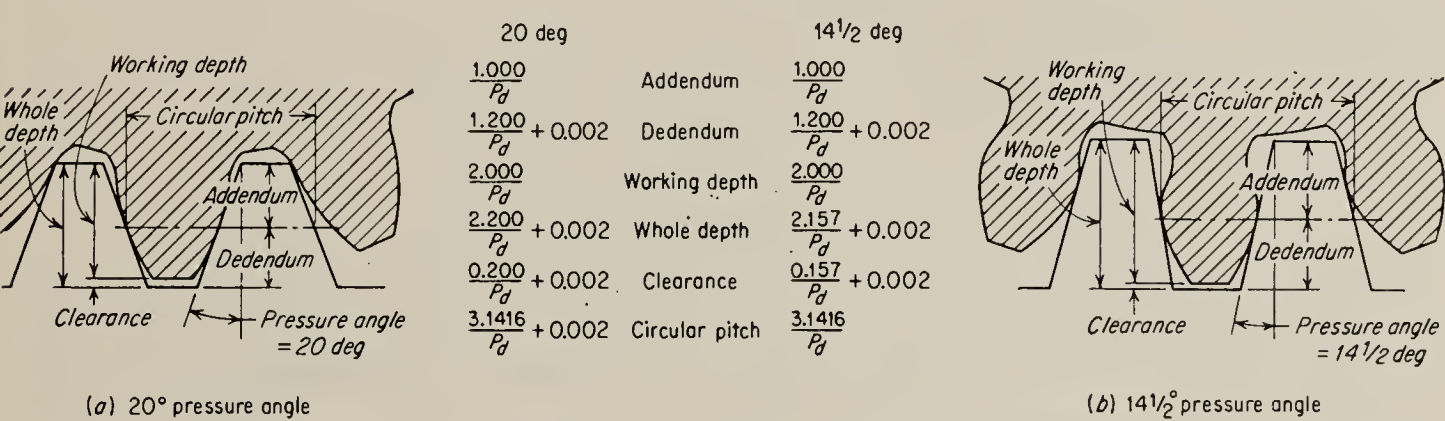


FIG. 8.13. Comparison of the standard pressure-angle designs for fine-pitch service. (Martin)

that scoring of the gear teeth can be eliminated by increasing the pressure angle to 25°. Especially for small pinions the higher pressure ratio provides higher surface durability and greater beam strength. It is possible to design high-pressure-ratio gears with stubbed teeth to increase beam strength and to yield a more favorable contact ratio.

Contact ratio is defined as the length of the path of contact between two teeth divided by the tooth pitch (circular pitch), along the path of contact, and it gives the average number of teeth in contact at any time. For pinions and small gears the contact ratio for 20°-pressure-angle gears may be more than twice that of 14½°-pressure-angle gears.³ This allows use of a thinner gear or one made of lighter material. Furthermore, owing to the undercut pinions in 14½°-gears, the gear action is not continuous, and the composite error is greater.⁴ The 20°-involute fine-pitch system is the present American Gear Manufacturers' Association (AGMA)

¹ L. D. Martin, Instrument Gears, *Machine Design*, vol. 26, no. 2, p. 129, February, 1954.

² *Ibid.*

³ Davison, Practical Considerations in Instrument Gear Design, *Product Eng.*, vol. 22, no. 9, pp. 183–187, 1951.

⁴ *Ibid.*

standard, and indications are that even higher pressure angles may be adopted in the future.

Gears are usually manufactured by cutting or hobbing the teeth in the blank with special machine tools. With small pinion gears there are two other methods that may also be used. Small pinions may be extruded, or they may be cold-drawn. Only the softer nonferrous metals such as bronze and aluminum lend themselves to extrusion, but any metal with good cold-working properties may be cold-drawn.

Cold-drawing has several advantages over the two other methods. The working strength of the teeth is increased. The method produces a smooth, hardened surface, and the dimensional accuracy of the gear is greater than with the hot extrusion process. Surface hardness up to 50 per cent greater and durability of 200 per cent more than that obtained in the other processes are possible.¹

The *helical* gear (Fig. 8.12*b*) and the *herringbone* gear are modifications of the spur gear. In the helical gear the gear teeth are slanted or helical with respect to the gear disk. The helical gear is a smoother-running gear than the spur gear, since the entire tooth does not make contact at the same instant with its mate. A disadvantage of the helical gear is the side thrust due to pressure of the mating teeth. This may be overcome by combining a right-handed and a left-handed helical gear on the same shaft. If the two gears are joined as a unit, the result is called a herringbone gear. The herringbone gear is used only in very high power applications such as nautical propulsion units, and is of little interest, therefore, for control systems.

Both the spur gear and the helical gear may be convoluted into various shapes. If the large gear is laid out flat, the result is called a *rack and pinion*. If the gear shafts are not held parallel but placed at an angle, the gear teeth must be modified, and the result is a *bevel gear and pinion* (see Fig. 8.12*c*).

The *pitch* of a gear is defined as the number of teeth on a disk with a 1-in. diameter and is thus a measure of the fineness of the gear teeth. For fractional-horsepower applications the standard production gear pitches are 32, 48, and 64. The 48 pitch is recommended for standard applications. For applications in which extreme smoothness is desired, the 64 pitch should be used. Where smoothness is not important and long gear life under heavy loads is the prime consideration, the 32 pitch may be used.

A gear that is somewhat different from those described above is the *worm and gear* (see Fig. 8.12*d*). The worm is a form of screw thread which bears on a large gear set with its axis at right angles to the axis of the worm. The gear is turned by rotation of the worm. The main

¹ E. H. Rathbone, Advantages of Cold-drawn Pinions, *Product Eng.*, vol. 21, no. 12, p. 114, 1950.

advantage of the worm-gear drive is that ratios of 100:1 or more may be obtained with one set of gears. Usually worm gears are *self-locking*; i.e., the gear cannot drive the worm. In some applications this is desirable, but in a servo control system the nonlinear action may cause stability problems. It can be shown, however, that worm gearing can be made fully as efficient as spur gearing¹ and thus may be made reversible.

Self-locking of a worm-gear drive occurs under conditions of heavy friction or small lead angle. The efficiency of a worm and gear may be

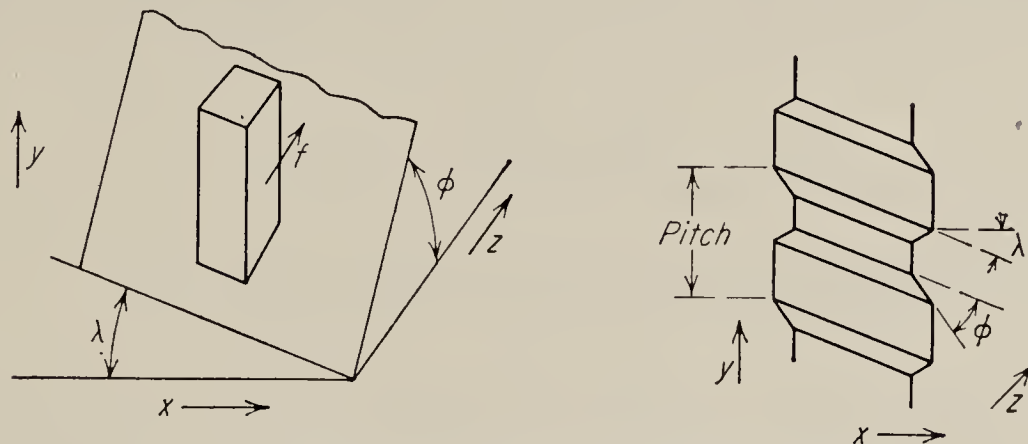


FIG. 8.14. Worm gearing with angles defined.

derived by considering the wedge, as shown in Fig. 8.14. The forces acting on the wedge may be expressed as components.

$$\mathcal{F}_x = \mathcal{F} \cos \phi \sin \lambda + f \mathcal{F} \cos \lambda \quad (8.9)$$

$$\mathcal{F}_y = \mathcal{F} \cos \phi \cos \lambda - f \mathcal{F} \sin \lambda \quad (8.10)$$

Here \mathcal{F} is the total force between the bodies, λ is the lead angle of the worm, at the pitch diameter, ϕ is the pressure angle, or the slope of the tooth, and f is the coefficient of friction. There is also a component of force in the z direction which tends to move the worm and gear apart, but since this motion is restrained by the shafts, it does not enter into the efficiency calculation. Solving Eqs. (8.9) and (8.10) simultaneously gives

$$\mathcal{F}_x = \mathcal{F}_y \frac{\cos \phi \sin \lambda + f \cos \lambda}{\cos \phi \cos \lambda - f \sin \lambda} \quad (8.11)$$

In an ideal gear train the coefficient of friction is zero, and Eq. (8.11) reduces simply to

$$\mathcal{F}_x = \mathcal{F}_y \tan \lambda \quad (8.12)$$

We shall define the efficiency of the gear train as the ratio of Eq. (8.12) to Eq. (8.11). The result is

$$\eta = \frac{\tan \lambda (\cos \phi \cos \lambda - f \sin \lambda)}{\cos \phi \sin \lambda + f \cos \lambda} \quad (8.13)$$

¹ Berard, Waters, and Phelps, *op. cit.*

Equation (8.13) can be differentiated with respect to λ and set equal to zero to find the value of λ for maximum efficiency. The result is

$$\lambda = 45^\circ + \frac{1}{2} \tan^{-1} \frac{f}{\cos \phi} \quad (8.14)$$

The coefficient of friction, f , depends on the material and lubrication conditions; values of f are available in the handbooks. For typical cases the lead angle for maximum efficiency is in the neighborhood of 45° .

For the case of the gear driving the worm, Eqs. (8.9) and (8.10) are solved for F_y , and the coefficient f changes sign. The resultant efficiency is

$$\eta' = \frac{\cos \phi \sin \lambda - f \cos \lambda}{\tan \lambda (\cos \phi \cos \lambda + f \sin \lambda)} \quad (8.15)$$

If the worm is self-locking, η' will be less than zero, or

$$\tan \lambda < \frac{f}{\cos \phi} \quad (8.16)$$

Thus self-locking is dependent on friction and/or a small lead angle. Of course, a small lead angle is desirable since it makes the installation more compact and permits a higher over-all ratio. The high ratio is one of the principal reasons for choosing a worm and gear in the first place. Nichols¹ states empirically that self-locking gears should be avoided in closed-loop systems because of their unstabilizing influence. However, the objection is not to the irreversibility but rather to the possibility of a load or accelerating torque wedging the gears so that the worm could not turn when driven in the normal manner. Nonlinear oscillation can occur if the gears wedge. The drive motor may exert enough torque to break the gears free, but the load then is accelerated and may overshoot. If this overshooting occurs, the gears will wedge again and the oscillation will continue.

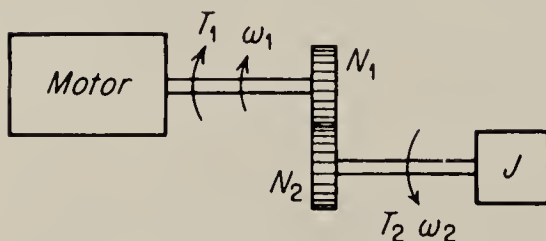


FIG. 8.15. A motor driving an inertia load through a set of gears.

8.7. Design of Gear Trains for Minimum Inertia. The moment of inertia placed on the shaft of a motor is a factor in its time constant. We now wish to determine how a gear ratio affects the load moment of inertia reflected to the motor shaft. It will then be possible to adjust the gear ratios to minimize this time constant. Figure 8.15

shows the configuration under consideration. Two relations may be written for the set of gears. First from the power consideration, in a

¹ James, Nichols, and Phillips, "Theory of Servomechanisms," Radiation Laboratory Series, vol. 25, McGraw-Hill Book Company, Inc., New York, 1947, p. 130.

perfect set of gears,

$$T_1\omega_1 = T_2\omega_2 \quad (8.17)$$

where T is torque and ω is speed. The second relation is that the speeds are inversely proportional to the gear ratio,

$$\frac{\omega_1}{\omega_2} = \frac{N_2}{N_1} \quad (8.18)$$

The Newton's law relation may be written for the output shaft load as

$$J = \frac{\hat{T}_2}{\hat{\alpha}_2} = \frac{\hat{T}_2}{s\hat{\omega}_2} \quad (8.19)$$

Substituting for the torque and speed from Eqs. (8.17) and (8.18), we have

$$J = \frac{\hat{T}_1\hat{\omega}_1/\hat{\omega}_2}{s\hat{\omega}_1N_1/N_2} = \frac{\hat{T}_1N_2/N_1}{s\hat{\omega}_1N_1/N_2} = \frac{\hat{T}_1}{s\hat{\omega}_1} \left(\frac{N_2}{N_1} \right)^2 \quad (8.20)$$

The quantity $\hat{T}_1/s\hat{\omega}_1$ represents a moment of inertia at the motor shaft, and it is related to the moment of inertia on the load shaft by

$$\frac{\hat{T}_1}{s\hat{\omega}_1} = J \left(\frac{N_1}{N_2} \right)^2 \quad (8.21)$$

Thus the load moment of inertia reflected to the motor shaft is reduced by the gear ratio squared. Therefore with only a moderate gear reduction the reflected load moment of inertia becomes negligible. The moment of the gears themselves is quite often not negligible, however.

In order to minimize the inertia of the gear train, the designer is tempted to place the maximum practical ratio in the first set of gears, thus making the effect of the following sets negligible. Let us examine the situation to see if this procedure is correct. Figure 8.16 shows the configuration under consideration. The total inertia seen at the motor shaft is

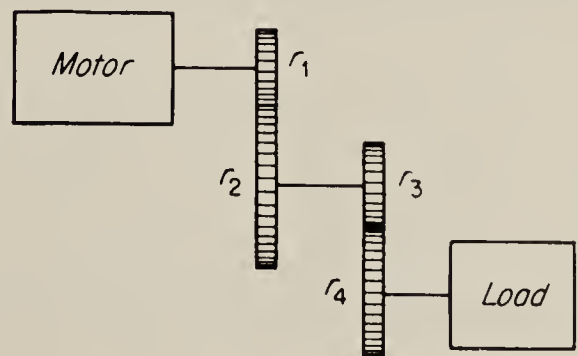


FIG. 8.16. Motor and a two-mesh gear train.

$$J_m = J_1 + (J_2 + J_3) \left(\frac{r_1}{r_2} \right)^2 + J_4 \left(\frac{r_1}{r_2} \right)^2 \left(\frac{r_3}{r_4} \right)^2 + J_L \left(\frac{r_1}{r_2} \right)^2 \left(\frac{r_3}{r_4} \right)^2 \quad (8.22)$$

where the J 's are the inertias of the gear disks and the r 's are the radii of the disks. We shall assume that the reflected load inertia is negligible; the inertia of r_4 would then also be negligible, but it must be included in the computation in order to solve for the individual gear ratios. Normalizing (8.22) to the inertia of the pinion J_1 and assuming that the

pinion inertias J_1 and J_3 and the radii r_1 and r_3 are identical,¹ we have

$$\frac{J_m}{J_1} = 1 + \left(\frac{J_2}{J_1} + 1 \right) \left(\frac{r_1}{r_2} \right)^2 + \frac{J_4}{J_1} \left(\frac{r_1^2}{r_2 r_4} \right)^2 \quad (8.23)$$

The moment of inertia is proportional to the fourth power of the disk radius if we assume that all the disks are of equal thickness. Hence we may write

$$\frac{J_m}{J_1} = 1 + \left(\frac{r_2^4}{r_1^4} + 1 \right) \frac{r_1^2}{r_2^2} + \frac{r_4^4}{r_1^4} \left(\frac{r_1^2}{r_2 r_4} \right)^2 \quad (8.24)$$

$$= 1 + \frac{r_2^2}{r_1^2} + \frac{r_1^2}{r_2^2} + \frac{r_4^2}{r_2^2} \quad (8.25)$$

For a given over-all ratio, $r_1 r_3 / r_2 r_4$ is constant. Thus, since r_1 and r_3 are constant, $r_2 r_4$ is constant:

$$(r_2 r_4)^2 = \left(\frac{r_1 r_3}{N_{\text{over-all}}} \right)^2 = K_1$$

Letting

$$r_1^2 = r_3^2 = K_2$$

we have

$$\frac{J_m}{J_1} = 1 + \frac{r_2^2}{K_2} + \frac{K_2}{r_2^2} + \frac{K_1}{r_2^4} \quad (8.26)$$

Differentiating Eq. (8.26) with respect to r_2 and setting to zero will give the r_2 for minimum inertia:

$$\frac{d}{dr_2} \left(\frac{J_m}{J_1} \right) = \frac{2r_2}{K_2} - \frac{2K_2}{r_2^3} - \frac{4K_1}{r_2^5} = 0 \quad (8.27)$$

$$r_2^6 - K_2^2 r_2^2 - 2K_1 K_2 = 0 \quad (8.28)$$

As an example, take the over-all ratio as 10 and let the pinions have a pitch diameter of $\frac{1}{2}$ in. Then

$$K_2 = (0.25)^2 = 0.0625$$

and

$$K_1 = \left[\frac{(0.25)(0.25)}{1/10} \right]^2 = 0.39$$

As an approximation, neglect the r_2^2 term, which will be about one-tenth the constant term and will make only a 1 per cent change in r_2 . Then

$$r_2 = [2(0.0625)(0.39)]^{1/6} = (0.049)^{1/6} = 0.605$$

giving a ratio of $r_2/r_1 = 0.605/0.25 = 2.42$. The second mesh would then make up the rest of the over-all ratio, 4.14 in this case. This

¹ The pinions are chosen as small as practical and in instrument gearing are usually identical. The more general case of power gearing, in which load capacity must be included, is considered below.

arrangement results in a reduction of inertia by a factor of 10 over the case of taking the whole 10:1 ratio in one step.

Figure 8.17 is a plot which shows the reflected inertia for any ratio for up to five meshes. Figure 8.18 is a nomograph which solves the equations given above for any over-all ratio. It may be seen from Fig. 8.17 that the reduction of inertia is relatively slight for four and five meshes

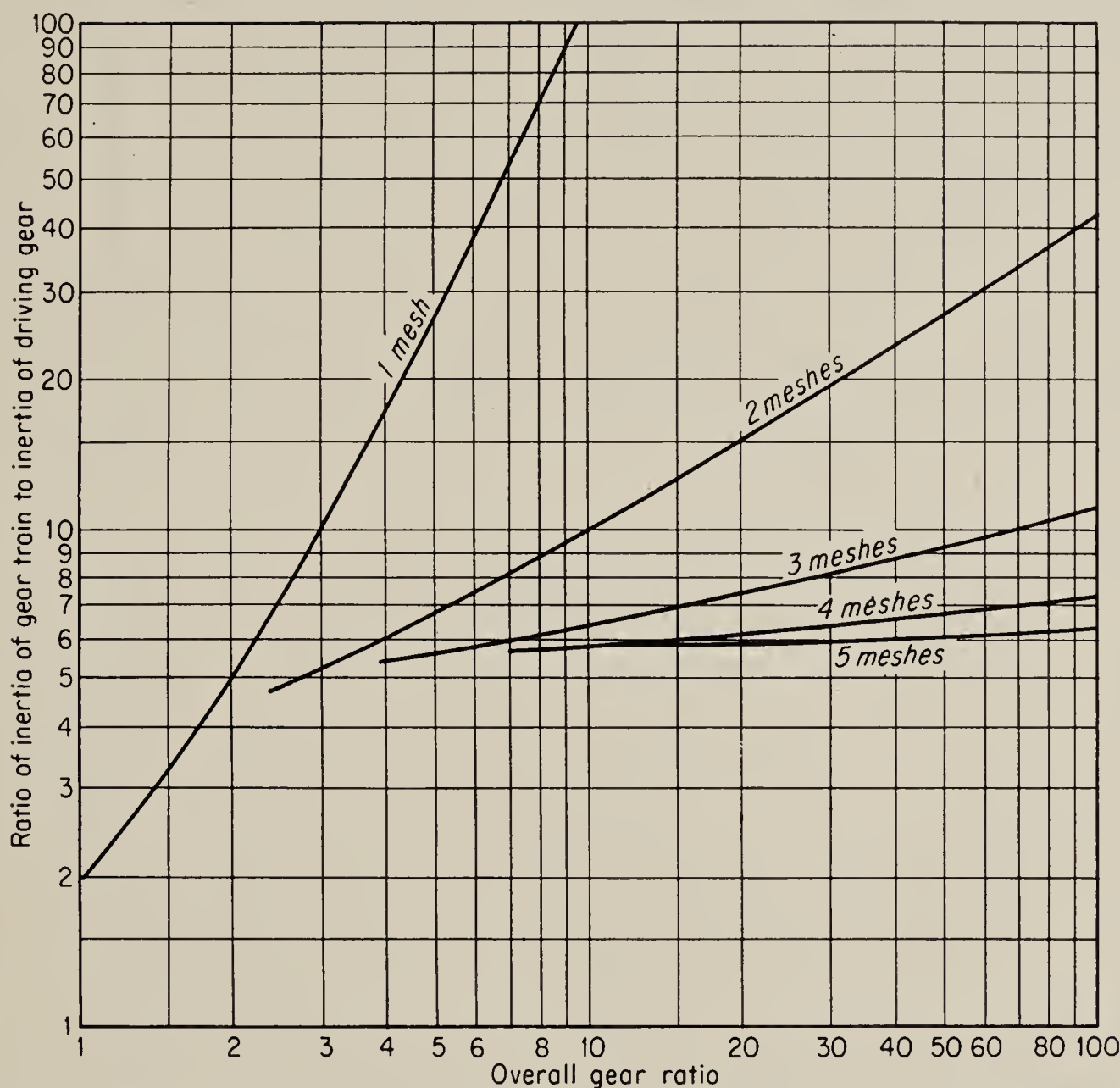


FIG. 8.17. The reflected inertia as a function of over-all gear ratio with the number of meshes as a parameter. The multimesh curves are computed for optimum apportioning of the ratio between the meshes. Note the relatively small improvement of four and five meshes over three meshes. (Courtesy Reeves Instrument Corp.)

compared with three meshes. For any over-all ratio up to 100 it is probably desirable to use a gear train of not more than three meshes. The number of meshes should be kept at a minimum to minimize the effect of backlash (Sec. 8.9).

In power gear trains, as differentiated from instrument gear trains, the design of the gear train is influenced by the strength of the materials used. If the power transmitted through a gear train is assumed constant, the torque increases in direct proportion to the speed reduction or gear-ratio

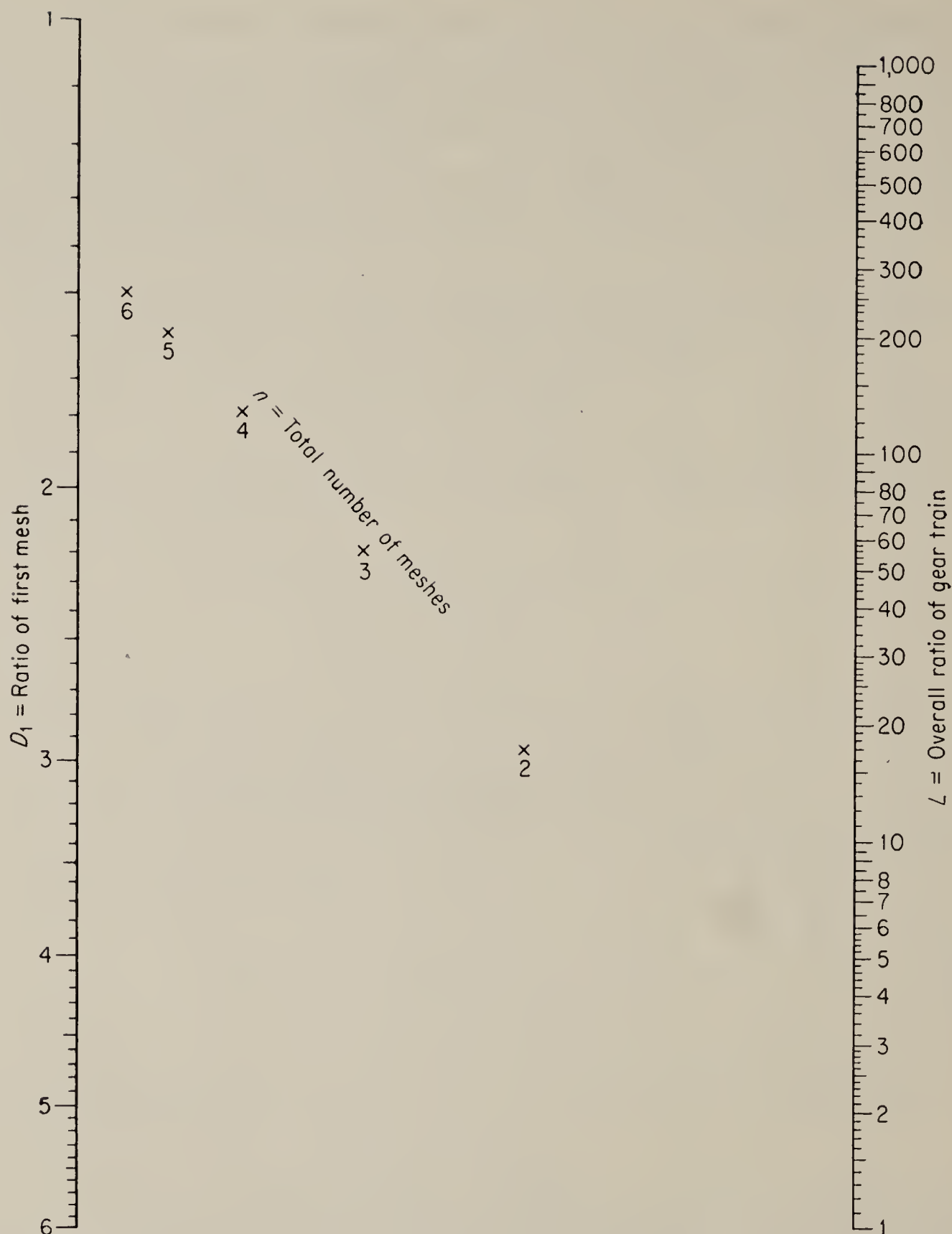


FIG. 8.18. A nomograph that solves the equations given in the text. Place a straightedge on the right-hand scale at the over-all gear ratio. The straightedge must also lie on the point giving the number of meshes. The straightedge then intersects the left-hand scale at the optimum ratio for the first mesh. The process may be repeated for the subsequent meshes. The example in the text may be checked using these curves. (Courtesy Reeves Instrument Corp.)

reduction. Thus the pinion gears in the second and following meshes must be made wider in order to support the increased forces. It will be shown that the calculation for power gearing can be significantly changed when this requirement is included.

There are two standard empirical relations that are used to calculate

the required size of power gears, the Lewis beam-strength formula and the Buckingham wear-load formula. In a first approximation¹ both of these relations can be reduced to

$$WD_p^2 = Km \quad (8.29)$$

where W is the face width of the meshing gears, D_p is the pitch diameter of the driving pinion, m is the gear ratio from motor to given pinion, and K is a constant of proportionality. Thus it is seen that the face width and/or the pitch diameter of the pinions must be increased as we progress through the gear train. This increase must be made when the forces in the gear train are an appreciable portion of the maximum allowable force for the material and gear design used. When over-all gear ratios are greater than 100 or when four or more pinions are used, this increase in pinion size will result in a considerable change in the gear design.

From practical fabrication and utilization considerations, Peterson² suggests that face width and pitch diameter of successive pinions be chosen by the empiric relation

$$\frac{W}{W_0} = \frac{D_p}{D_{p0}} = \frac{P_0}{P} = m^{1/3} \quad (8.30)$$

where P is the diametral pitch of the gear and the subscript 0 refers to the motor pinion.

The solution for minimum inertia employing the criterion in Eq. (8.30) is displayed graphically for various meshes in Fig. 8.19. It will be noted that, for small over-all ratio and low number of total meshes, the results from Fig. 8.19 are identical with those from Fig. 8.17, which was computed for constant-sized pinions. For higher ratios, however, the use of Fig. 8.19 will result in a reduction in inertia of greater than 2 to 1.

The advantages of constant-sized pinions are, of course, all the advantages of standardization. It should be borne in mind also that, where the force transmitted is negligible with respect to the strength of the gears, all the gears may be reduced in thickness and Fig. 8.19 will still yield optimum results. Figure 8.20 gives curves that allow calculation of the individual gear meshes after the total number of meshes has been chosen from Fig. 8.19.

A possible compromise between the two extremes of complete standardization and individual design exists for medium- and light-duty gear trains. It is possible at the slow-speed end of the train to use gears of lower pitch than those used at the high-speed end. The low-pitch gear teeth are larger and stronger and are more suited for heavy duty. The over-all cost of the gear train is kept low by employing standard gear

¹ D. Peterson, Power Gear Trains, *Machine Design*, vol. 26, no. 6, p. 161, 1954.

² *Ibid.*

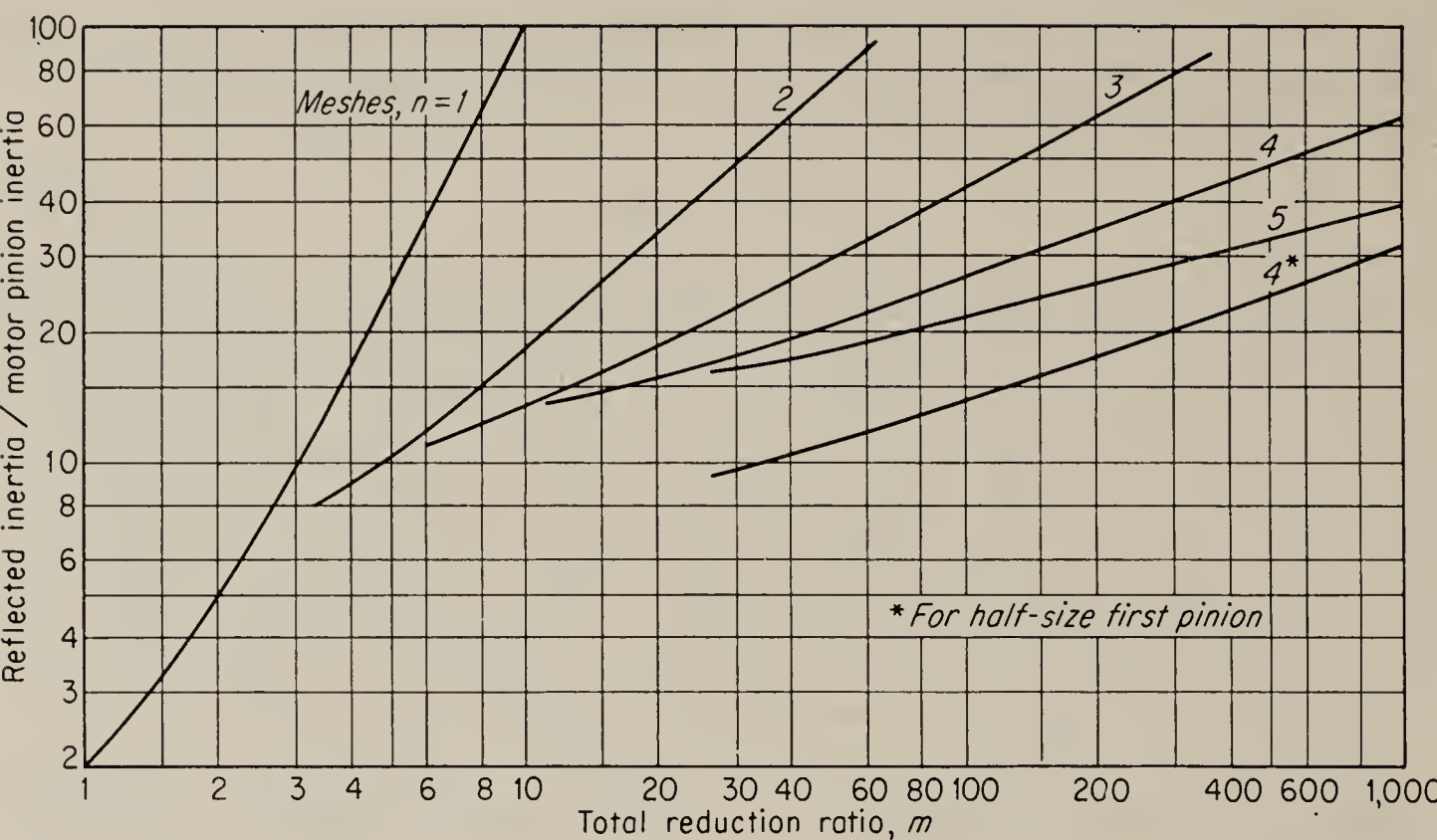


FIG. 8.19. Reflected inertia as a function of gear meshes for power gears. (From Peterson)

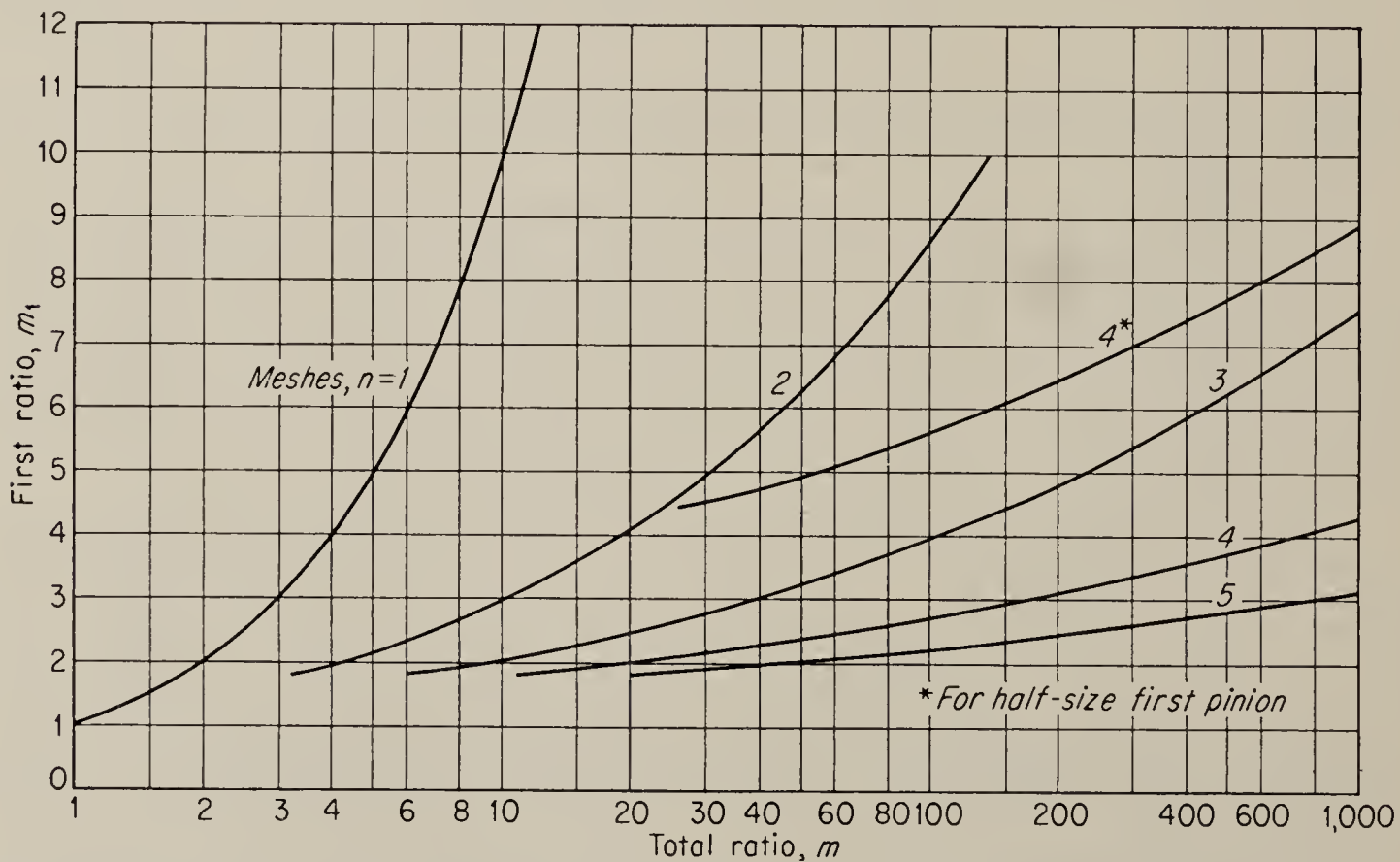


FIG. 8.20. Choice of individual meshes given over-all ratio and number of meshes. Subsequent meshes are chosen by reducing ratio and number of meshes by the first mesh choice. (From Peterson)

itches; yet the strength requirements are met. This solution is widely used in practice.

Quite often, the inertia of the gear train can be reduced by measures other than optimizing the ratios. By using punched gear disks rather than solid disks; the inertia can be reduced to about 75 per cent of the

solid-disk value. A *punched gear disk* is one in which much of the material has been removed from the otherwise solid body of the disk. Substitution of materials can work an even greater reduction. The substitution of aluminum gears for steel reduces the inertia to 33 per cent of its original value, and the substitution of aluminum for brass gears reduces the inertia to 25 per cent of its original value. Of course, factors such as tensile strength and durability must be considered when substituting materials.

The use of nylon for medium- and light-duty gears is also a possibility. The extreme lightness of nylon (specific gravity, 1.14) is its main attraction, although it is also easy to fabricate, resilient, and corrosion-resistant. Nylon has a low coefficient of dry friction, which in effect makes it self-lubricating, and it resists wear and abrasion. Nylon's ultimate tensile strength is about 12,000 psi, and it should not be subjected to environmental temperatures of greater than 100 to 120°F because it rapidly loses its tensile strength and form at high temperatures.¹

8.8. Gear Ratio for Load Matching. In instrument gearing the load is usually negligible, and the inertia of the gear train is the only factor that must be considered. The gear ratio is chosen on the basis of accuracy and/or loop gain. The motor is chosen for its time constant, including the inertia of the gear train.

In power applications, however, the power requirements of the load and the power capacity of the motor become important. It is usually necessary to choose the smallest motor that is capable of supplying the load in order to minimize size, weight, and cost of the power package. The over-all gear ratio will thus be chosen with this in mind. The number of meshes and ratios of each mesh should be designed by the methods discussed in Sec. 8.7 for power gear trains.

In Fig. 8.21 is given, as an example, the approximate torque and horsepower plotted against speed for a typical a-c servomotor. In this particular case, the motor develops maximum power at $N_s/2$, but whatever the shape of the motor speed-torque curve, the speed for maximum power may be found. Assuming that the required load velocity is known, the over-all gear ratio should be chosen so that the motor is operating at the speed at which it develops maximum horsepower when the load is oper-

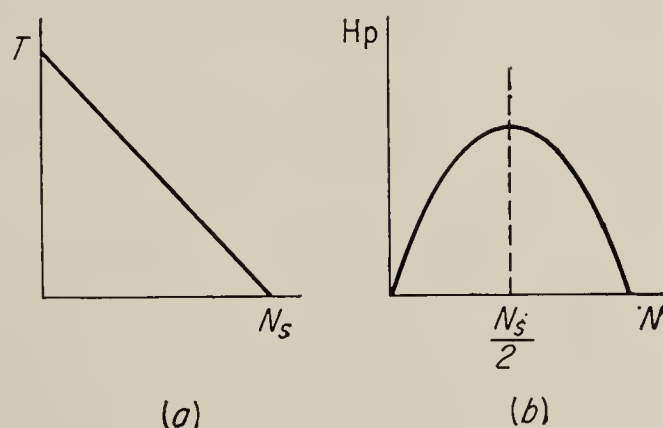


FIG. 8.21. (a) Torque versus speed and (b) horsepower versus speed of typical a-c servomotor.

¹ R. Zimmerli, Designing Fabricated Nylon Parts, *Machine Design*, vol. 26, no. 3, March, 1954, pp. 153-159.

ating at its design velocity. The horsepower rating of the motor can be determined by determining the speed-torque characteristic of the load and reflecting it through the gear train to the motor shaft. For instance, in Fig. 8.22 are shown the horsepower and torque curves for a load consisting of viscous damping and inertia.

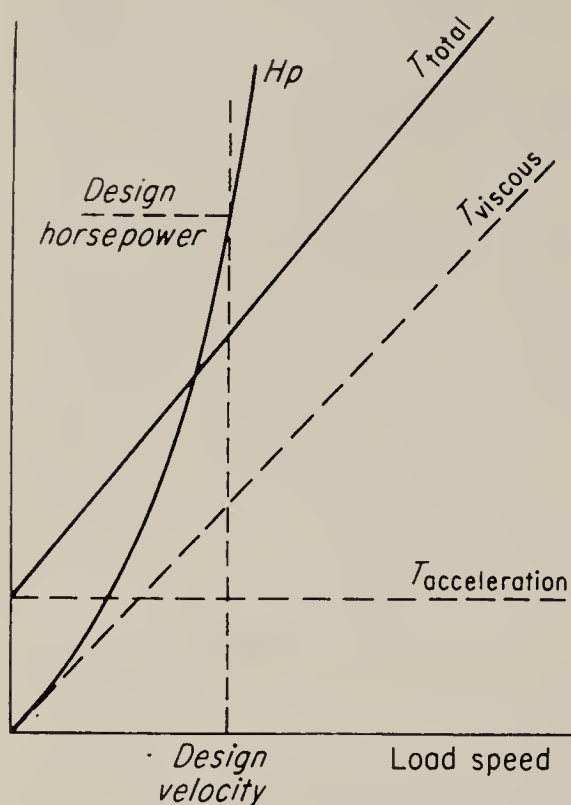


FIG. 8.22. Horsepower and T versus N for typical load.

The maximum velocity of the load and the maximum acceleration are specified. When the reflected load horsepower-speed curve and the torque-speed curve are superimposed on the motor curves, the motor output must exceed the load by a given design factor, usually 1.5 or 2 to 1.

8.9. Backlash in Gears. Backlash in gears is the looseness or play between the input and the output of the train. Backlash may be measured as the angular displacement through which the input can be moved with the output fixed, or it may be given as the linear distance along the pitch circle through which a gear may move with respect to its fixed mating gear. AGMA standards are given in inches and can be

converted to angles if the pitch diameter is known. The AGMA standard backlash classifications are shown in Table 8.2, along with composite error limits.

The *composite error* referred to in Table 8.2 is an effect which may be considered separately from backlash. In the cutting of the gear teeth and the locating of the center bore for the gear shaft, there will inevitably develop a certain amount of eccentricity between the shaft and the pitch circle as well as distortions of the pitch circle itself. The eccentricity that these errors lend to the gear-tooth motion as the gear is turned is called composite error. The distance between the extremes of this eccentricity is the total composite error, while the variation over the angle covered by one tooth is called the tooth-to-tooth composite error.

Backlash results in a nonlinear relation between input and output of the gear train and has received considerable study in the past few years.¹ For large-amplitude signals, backlash has little effect, but as the signal amplitude is reduced, backlash becomes a more important factor.

If a frequency-response characteristic of a portion of a system is to be taken, the standard procedure is to maintain a constant input magnitude or even reduce the magnitude as the frequency is increased. Let us

¹ Chestnut and Mayer, "Servomechanisms and Regulating System Design," John Wiley & Sons, Inc., New York, 1955, vol. II, sec. 8.1.

assume that the system contains a gear train with some backlash. Since, in most systems, the response drops off at high frequencies, the amplitude of the motion of the input gear will be reduced as the frequency is increased. On account of backlash, the output of the gear train will be reduced over and above the reduction that linear theory would predict at

TABLE 8.2. AGMA STANDARD 236.04 FOR ALLOWABLE BACKLASH
AND COMPOSITE ERROR IN GEARS
Standard Specified Backlash

<i>Diametral pitch</i>	<i>Backlash,* in.</i>
Class A	
20-45.....	0.004 -0.006
46-70.....	0.003 -0.005
71-90.....	0.002 -0.0035
Class B	
20-60.....	0.002 -0.004
61-120.....	0.0015-0.003
121 and finer.....	0.001 -0.002
Class C	
20-60.....	0.001 -0.002
61-120.....	0.0007-0.0015
121 and finer.....	0.0005-0.001
Class D	
Any pitch.....	No measurable backlash

Total and Tooth-to-tooth Composite Error Limits
for Spur, Helical, Worm, and Bevel Gearing

Class	Total composite error, in.	Tooth-to-tooth composite error, in.
Commercial 1.....	0.006	0.002
Commercial 2.....	0.004	0.0015
Commercial 3.....	0.002	0.001
Commercial 4.....	0.0015	0.0007
Precision 1.....	0.001	0.0004
Precision 2.....	0.0005	0.0003
Precision 3.....	0.00025	0.0002

* Between two assembled gears at their tightest point of mesh. Backlash will be increased when the low points of runout are in contact.

the high frequencies. Figure 8.23 shows an example of this phenomenon. If a constant output magnitude which is large enough to eliminate the effect of backlash is maintained throughout any tests, the effect of backlash will not appear. The effect of backlash may be important to the stability of a closed-loop system, especially if it is a high-performance system with a rather narrow range of stability.

Backlash can be reduced by several means. One method is to use spring-loaded gears. A spring-loaded gear is a set of two identical spur gears placed side by side, as shown in Fig. 8.24. One of the pair is fixed to the shaft as usual, but the other gear is loose on the shaft. The loose gear is attached to the fixed gear by springs. Thus the two spur gears are set to bear on the pinion with spring tension, so that the gears are in contact regardless of the rotation; in this manner the backlash is reduced. Spring loading increases the rate of wear of the gears, because the gears are always fully loaded. Further, if the springs are not stiff enough, there will be some "give" when the train is reversed. This may lead to

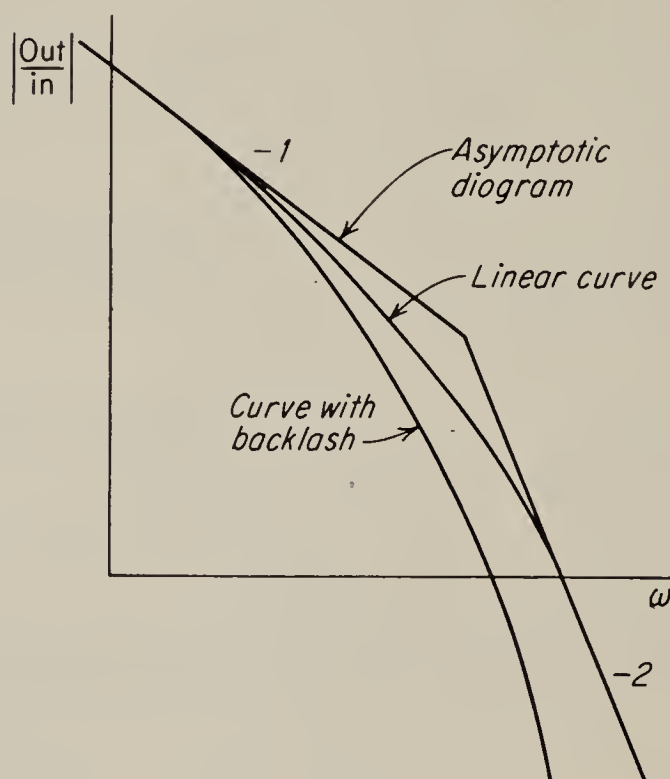


FIG. 8.23. Distortion of experimental frequency response due to backlash in a gear train.

instability. Finally the double gear increases the train inertia and requires a longer pinion.¹

A second possibility is to reduce the number of gears in the train. If the over-all ratio is fixed, this requires a larger ratio in the first stages than would be dictated by minimum-inertia considerations. The final choice of the number of gear meshes is thus a compromise between reduction of the effect of backlash and reduction of inertia. The reduction of inertia must usually, therefore, be achieved by other means than increasing the number of gear meshes.

Another possibility for reducing backlash is to use a gear train with a very fine pitch and to set the shafts so that the gears bear rather tightly. This increases bearing and rubbing friction, however, and thus cannot be carried too far. While it is theoretically possible to adjust the operating

¹ G. W. Michalec, Precision Gearing, *Machine Design*, vol. 27, p. 202, February, 1955.

centers of conventional circular-involute gears and still obtain proper gear action, arrangements to force the gear centers together by screw adjustments on the bearings, etc., result in rapid wear and thus reduce the backlash only temporarily.

A final solution to the problem of backlash is to adopt an entirely different gear design based on the conical involute rather than the conventional circular involute. This design is called *tapered-tooth involute gearing*, or *beveloid gearing*. The cross section of a gear tooth developed

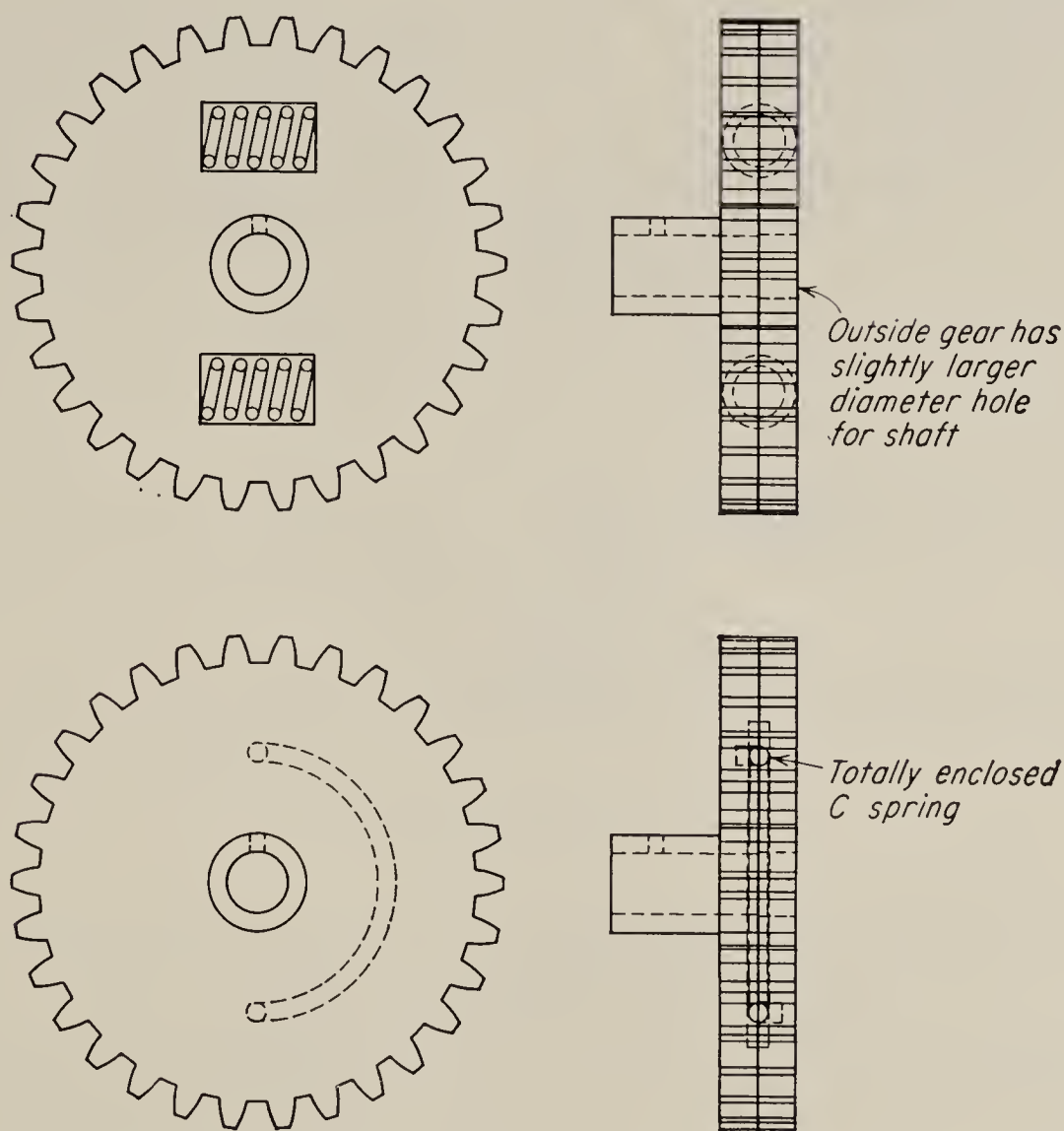


FIG. 8.24. Spring-loaded spur gears. Two forms of tension springs are shown.

by conventional design is constant throughout the tooth thickness, while the gear tooth resulting from the conical involute is tapered, hence the name. Figure 8.25 illustrates this. The conical-involute gear has tapered tooth thickness, tapered root, and, in most cases, tapered outside diameter.

The advantages of beveloid gearing¹ are ease in holding manufacturing accuracy, since precision tolerances in the range of 0.0002 in. total composite error are possible, unlimited meshing combinations, and freedom of gear arrangements. This last advantage means that beveloid gears are not designed for a given mounting angle or mounting center. Beveloid

¹ A. Beam, Beveloid Gearing, *Machine Design*, vol. 26, no. 12, p. 220, 1954.

gears allow adjustment of mounting angle, mounting distance, and direction of axes. This means that backlash can always be eliminated by sliding the gears on their shafts or by adjusting the shafts. Beveloid gears have two major disadvantages: First, they are somewhat more expensive to manufacture than conventional gears, since automatic machines are not used. Second, with gears with intersecting or skew shafts, the contact between gears is theoretically a point, a fact which

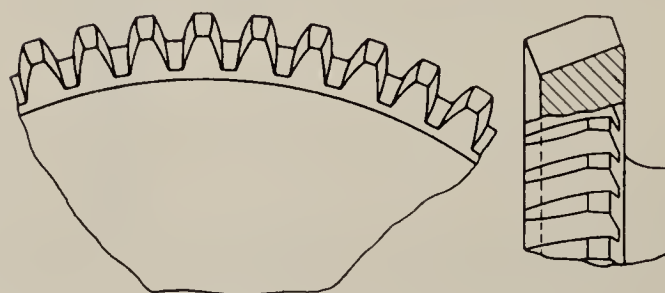
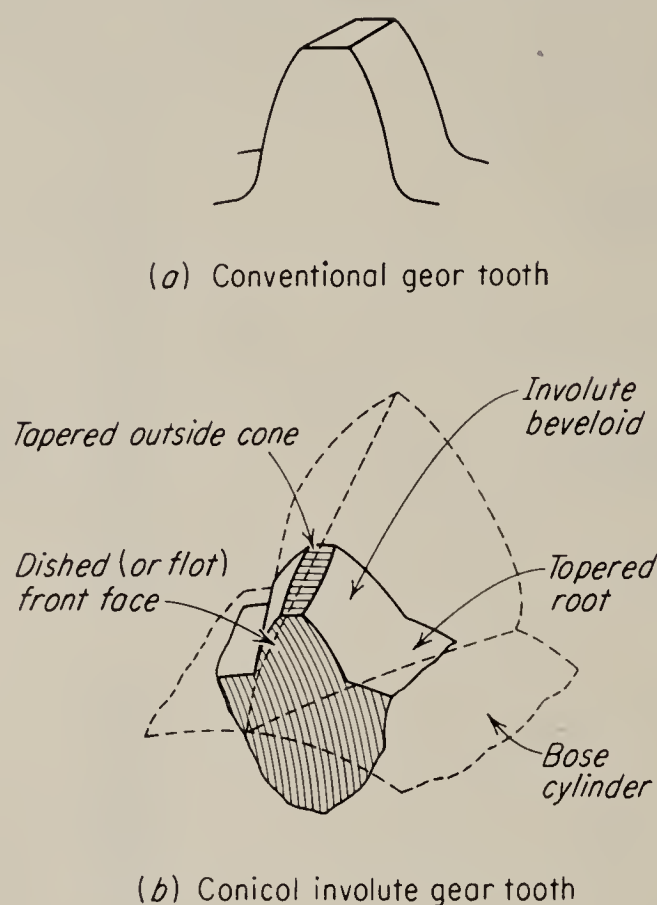


FIG. 8.25. Cross section of conical-involute gear tooth.

limits the allowable loads. When the shafts are parallel, however, beveloid gear teeth have line contact and are thus equivalent in wear and load-carrying ability to conventional gears.¹ For small skew angles, beveloid gears approach the loading potential of conventional gearing.

8.10. Ball-screw Actuator. Several possible ways of converting rotary motion into linear motion exist. The rack and pinion and the worm and pinion are two examples. Another example is the simple screw thread and nut. If the nut is prevented from turning as the screw is turned, the nut travels linearly along the screw. The ball-screw actuator differs from

¹ *Ibid.*

the nut and screw only in that the operating surfaces rest on ball bearings. Figure 8.26 shows a ball-screw actuator. As the actuator moves along the screw, the balls are left behind. The tube shown in the figure carries the balls to the front of the actuator, where they reenter it. The advantage of the ball-screw actuator over conventional threads is the greatly

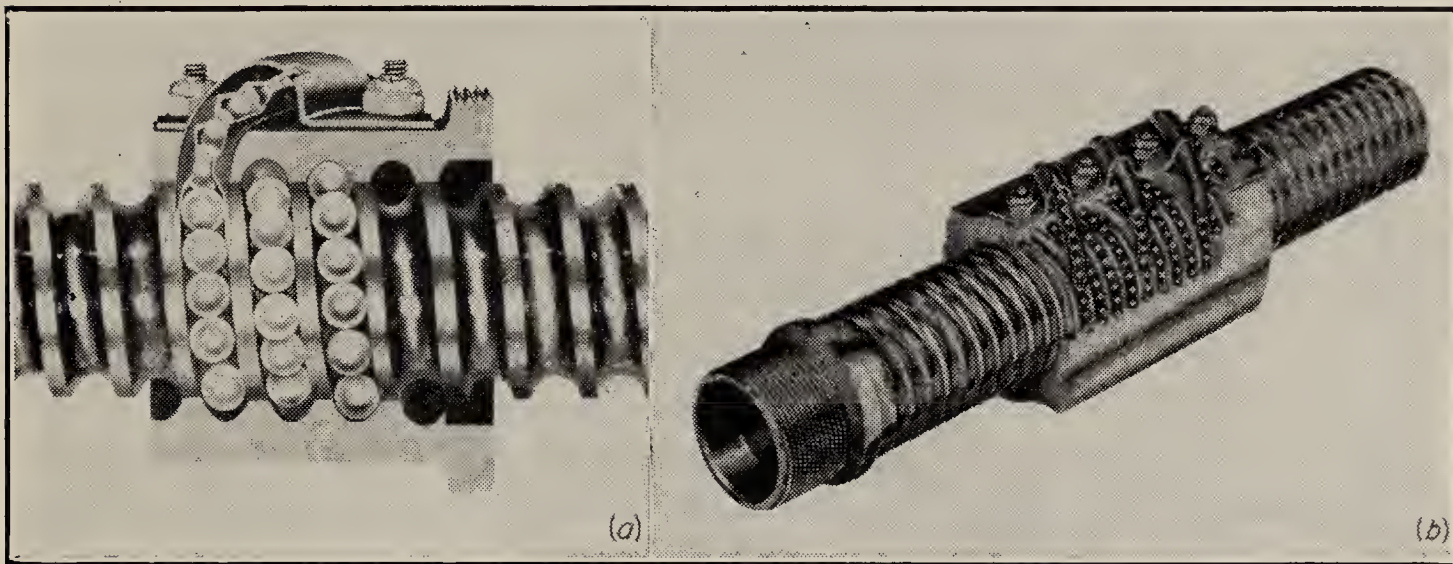


FIG. 8.26. The ball-screw actuator. (a) Typical ball-bearing screw assembly employing threaded nut with single ball circuit. Projecting guide finger on return tube deflects balls into tube for recirculation. (b) Heavy-duty ball-bearing screw assembly which has three ball circuits and yoke-type ball deflectors. (Courtesy Saginaw Steering Gear Division, General Motors Corp.)

increased efficiency of the former. Ball-bearing screws operate at efficiencies of over 90 per cent,¹ which is more than twice as efficient as a typical screw and nut.

PROBLEMS

8.1. Design an instrument gear train with an over-all reduction of 50 and three meshes. What are the relative decreases in inertia with two, three, four, and five meshes?

8.2. Design a power gear train with an over-all reduction of 50 and three meshes. What are the relative decreases in inertia with two, three, four, and five meshes? Compare these results with those in Prob. 8.1.

8.3. Consider the Lancaster damper shown in Fig. 8.9. Suppose that the inertia of the motor armature and the rigidly attached damping disk is J_1 , the inertia of the floating disk J_2 , and the coefficient of viscous friction B . The motor has a generated voltage coefficient k_v , a torque constant k_t , armature resistance R_a , and negligible armature inductance. (a) Find the transfer from \hat{E} to $\hat{\theta}$ without the damper. (b) Find the transfer from \hat{E} to $\hat{\theta}$ with the damper. (c) Letting $J_1 = 0.1$, $J_2 = 0.05$, $B = 0.5$, $R_a = 0.2$, $k_v = 0.3$, and $k_t = 0.3$ in a consistent set of units, plot the asymptotic α diagram for the device.

8.4. Find the transfer from E to θ for a tuned damper. Let the spring constant be K . Compare result to (b) in the previous problem.

¹ D. A. Galonska, Ball-bearing Screws, *Machine Design*, vol. 27, p. 201, October, 1955.

CHAPTER 9

MECHANICAL COMPONENTS

9.1. Introduction. In this chapter a number of mechanical components widely used in control systems will be considered. No attempt will be made to include material that is available in standard works on machine design and mechanisms; only the requirements that must be met by these components that are peculiar to their use in control systems will be discussed. The use of mechanical elements in computers has also been adequately discussed elsewhere¹ and will therefore not be considered here.

9.2. Differential Gears. Differential gears are used in control systems to add or subtract the positions or velocities of two shafts. The most common type is the *bevel-gear differential* shown in Fig. 9.1. Gear 1 is

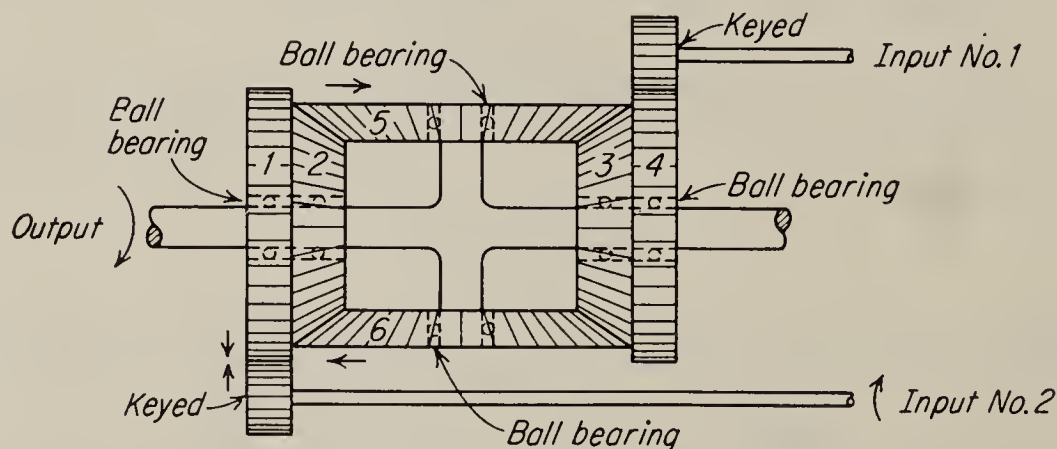


FIG. 9.1. Bevel-gear differential.

integral with gear 2, and gear 3 is integral with gear 4. Both of these two pairs of gears are arranged to rotate freely about the output shaft. The bevel gears 2 and 3 mesh with gears 5 and 6. Gears 5 and 6 also ride free of their shaft on ball bearings. Consider first that input 1 is fixed; then gears 3 and 4 cannot turn. If input 2 is turned in the direction of the arrow, gears 1 and 2 will also turn, as will gears 5 and 6, all in the arrow direction. Gears 5 and 6 will thus “walk” around on the stationary gear 3, and they will carry the output shaft with them, in the arrow direction. Now if input 2 is fixed and input 1 is turned, the output shaft will still turn in the direction of the arrow. When both input shafts are rotated

¹ Soroka, “Analog Methods in Computation and Simulation,” McGraw-Hill Book Company, Inc., New York, 1954, Chap. I. Svoboda, “Computing Mechanisms and Linkages,” McGraw-Hill Book Company, Inc., New York, 1948.

simultaneously, these two actions take place simultaneously, so that the output motion is the sum of the motions generated by input 1 and input 2, separately. Subtraction may be accomplished by defining one of the inputs in the opposite sense or by reversing its direction with a 1:1 spur gear. In a symmetrical differential gear train like that shown in Fig. 9.1, there is a 2:1 reduction from either input to output because of the “walking” action discussed above. The output velocity or position is therefore one-half the sum of the two input velocities or positions.

A common application of differential gears in a control system is as the device that subtracts the output of a servo from the input to yield the error. Differential gears have also been used in mechanical analogue

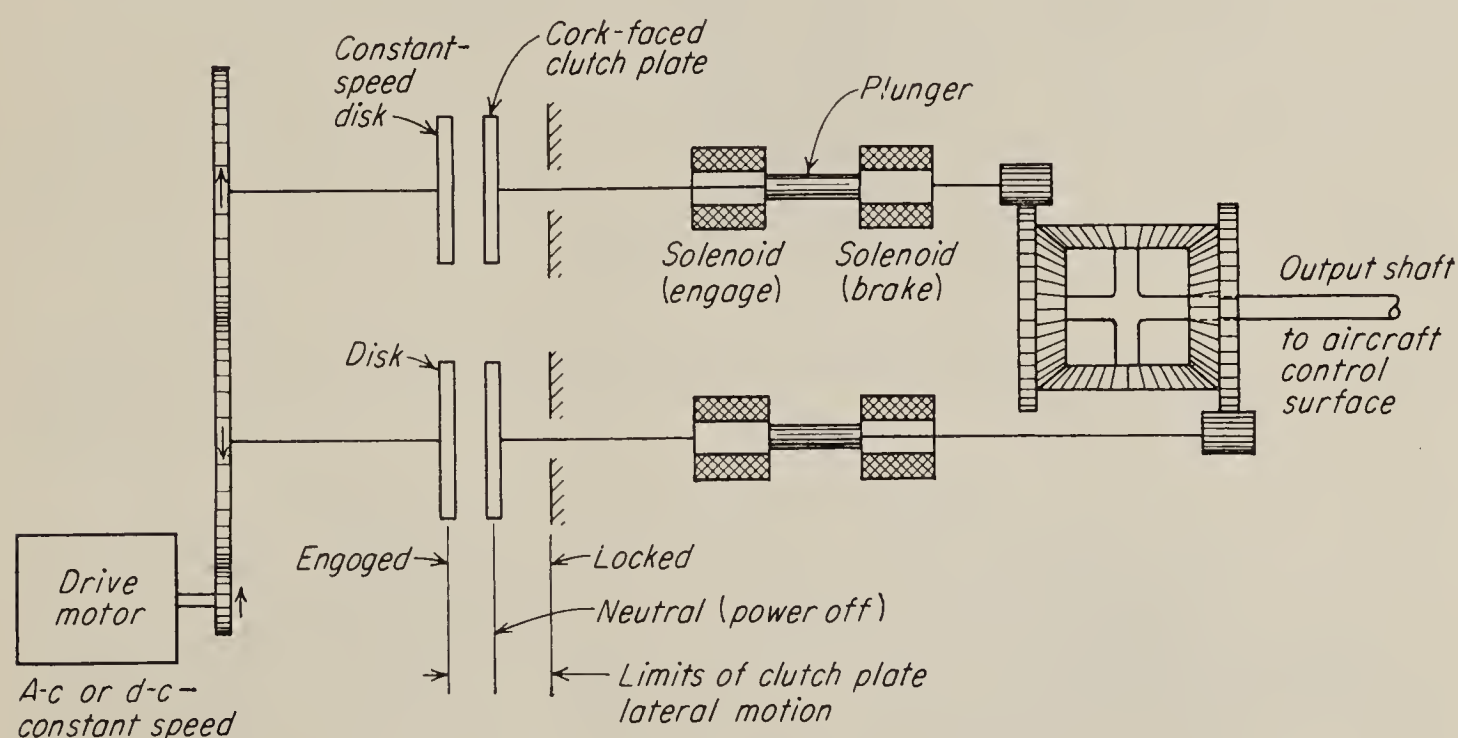


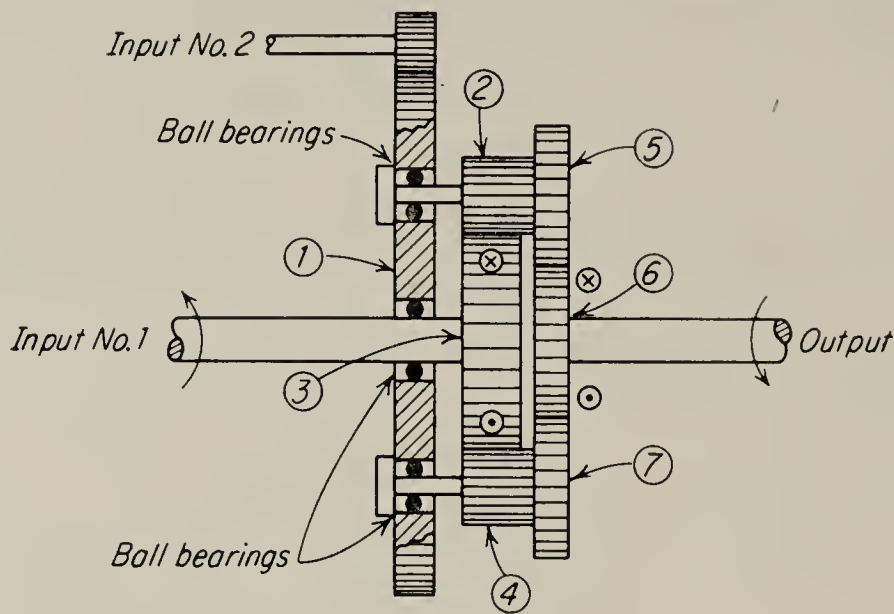
FIG. 9.2. On-off servo utilizing differential gear output.

computers as adders and subtractors. A somewhat different application is in the electromechanical clutch that was used as the power element of the C-1 autopilot. The unit is shown in a simplified schematic in Fig. 9.2. When power is applied to the energizing circuit, solenoids pull the clutches into contact with fixed bearing plates, thus locking the differential and the output shaft. When an input signal is fed into the component, one or the other of the clutch-plate brakes is released, and that plate is pulled into contact with the moving disk. The four solenoids are electrically interlocked so that only proper combinations may be energized at any one time.

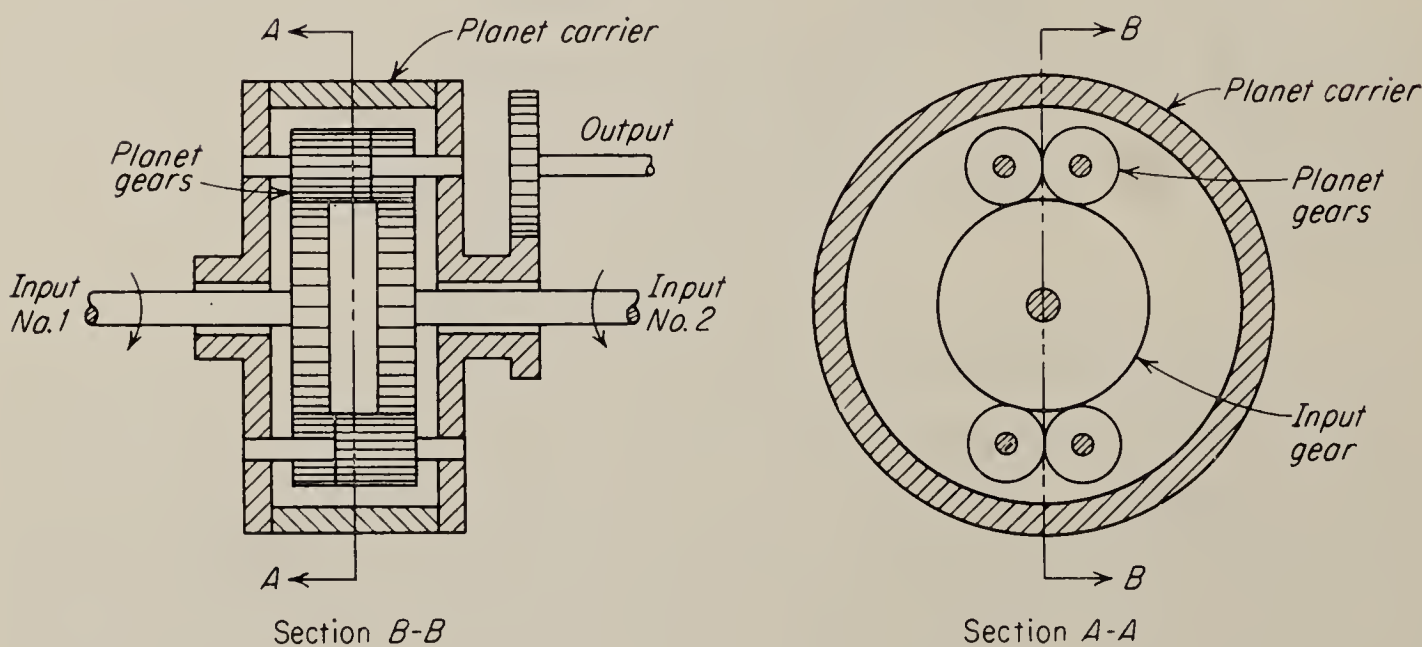
If the power fails or is turned off, the clutches return to the neutral position, and the controls are free so that the human pilot may manipulate them in a normal manner. Should one of the clutches fail to open during operation, the output shaft will still be free, and finally, if one clutch fails by sticking in either closed position, the shaft may be freed by turning the power off. The fact that with a differential gear both clutches must

operate for any force to be transmitted to the output shaft is considered a safety feature.

The *spur-gear differential* shown in Fig. 9.3*a* differs in several respects from the more common bevel-gear differential. It is quite compact, with a width of only three gears and the necessary supporting brackets.



(a) A minimum-width differential



(b) A device used in on aircraft-gun computer

FIG. 9.3. The spur-gear differential.

More important, the transfers from the two inputs to the output are not, in general, the same, and they are not independent of the gear ratios as in the bevel-gear differential.¹ Figure 9.3*b* shows a more symmetrical form of spur-gear differential used in an airborne computer. In Fig. 9.3*a* gear 3 is integral with input shaft 1. Gear pair 2 and 5 and gear pair 4 and 7 are integral. Let us assume that input 2 is fixed and that input 1 is moved

¹ Svoboda, *ibid.*

in the direction shown by the arrow. Gear 1 then cannot move, and gear 3 is moved by the input shaft in the direction shown. The two gear pairs are thus driven, and they in turn drive gear 6, which is integral with the output shaft. If input 1 is fixed, input 2 turns gear 1, which walks the gear pairs around the stationary gear 3. We note that the output shaft will only move if the ratio r_3/r_2 is not the same as r_5/r_6 . We have the further constraint that, if the input 1 shaft and the output shaft are to be collinear, as shown in Fig. 9.3a, it must be true that

$$r_3 + r_2 = r_5 + r_6 \quad (9.1)$$

From input 1 we have the relation

$$\frac{\text{Output}}{\text{Input}_1} = \frac{r_3 r_5}{r_2 r_6} \quad (9.2)$$

while from input 2, in terms of the radius of gear 1, we have

$$\frac{\text{Output}}{\text{Input}_2} = \frac{r_3}{r_2} - \frac{r_5}{r_6} \quad (9.3)$$

where the positive sense for input 2 is the same as for input 1. The net output is thus

$$\text{Output} = \left(\frac{r_3 r_5}{r_2 r_6} \right) (\text{input}_1) + \left(\frac{r_3}{r_2} - \frac{r_5}{r_6} \right) (\text{input}_2) \quad \text{radians} \quad (9.4)$$

Given either r_3/r_2 or r_5/r_6 , we may then select the other ratio to give any desired ratio of the two input-to-output ratios. While this flexibility is useful in adjusting gains, the most common arrangement is an input-output ratio of 0.5 for both inputs. In this case, r_3/r_2 and r_5/r_6 can equal 1.0 and 0.5, respectively.

9.3. The Universal Joint. The *Hooke joint*, or *universal joint*, is a device used to couple two shafts that meet at an angle. In Fig. 9.4 is shown a primitive form of the joint. The points A_1 and A_2 are the ends of the yoke on shaft A , and the points B_1 and B_2 are the ends of the yoke on shaft B . We are interested in the distortion of the angular motion produced by this device. Let us erect a plane normal to the shaft A and a plane normal to the shaft B , both passing through the point O . Then these two planes will meet at the angle α , the angle between the two shafts. The two planes will contain, respectively, the locus of the point A_1 and the locus of point B_1 . In Fig. 9.4d are represented the two planes and the loci of the two points A_1 and B_1 . If shaft A is moved through an angle θ , A_1 will move to A'_1 and B_1 will move to B'_1 . The angle A_1OB_1 and $A'_1OB'_1$ is always a right angle because it is held so by the physical cruciform. Furthermore the projection of $A'_1OB'_1$ on the plane normal to A is a right angle. This can be shown from the theorem of solid geometry

that the projected angle of two intersecting lines at right angles on a plane parallel to one of the lines is always a right angle. The plane normal to A is parallel to line OA_1 ; in fact it contains it. We have before us the problem of finding the angle ϕ , or $B_1OB'_1$, which the shaft B has actually turned through. The point β on the line of intersection of the two planes

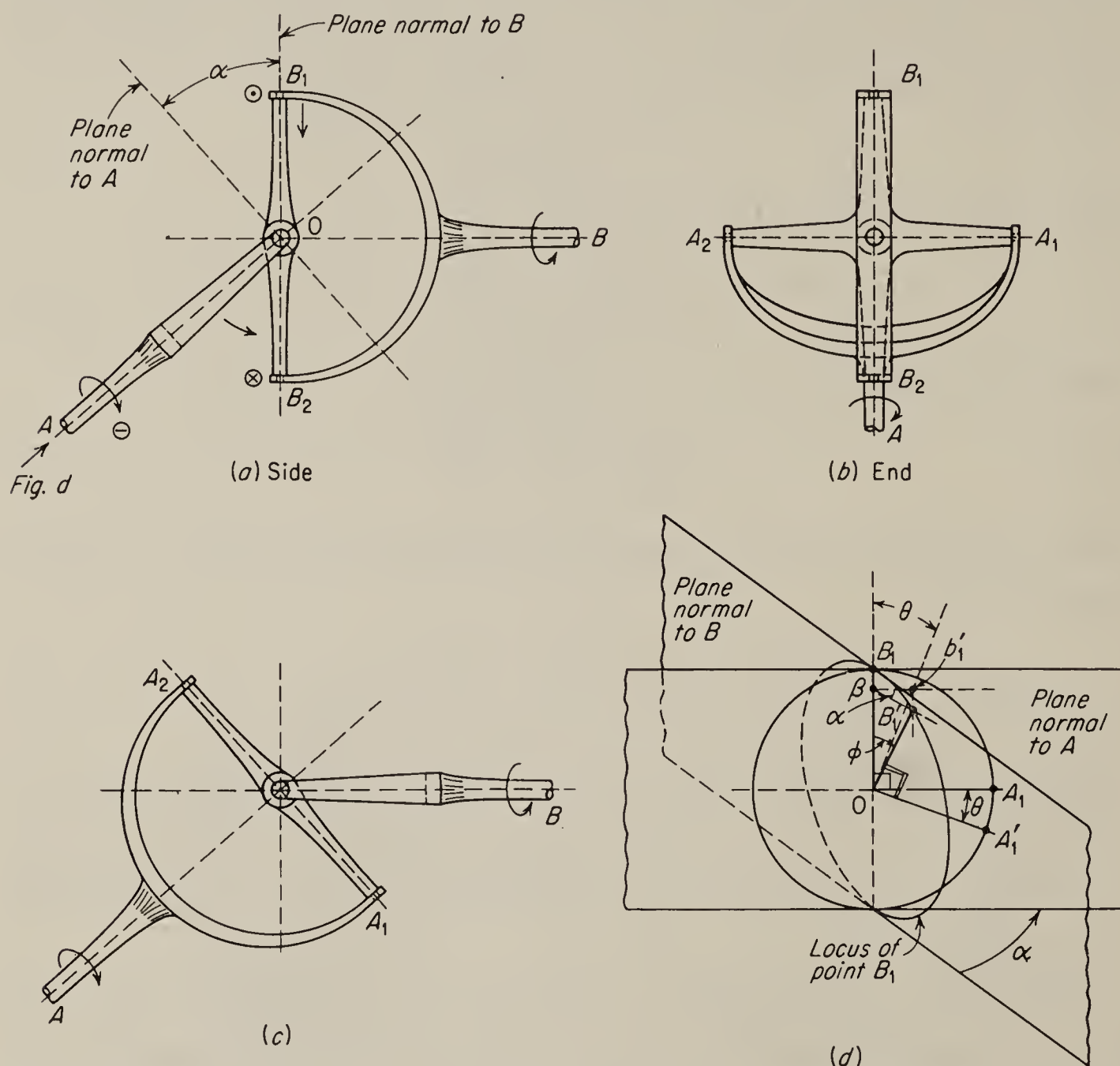


FIG. 9.4. The universal joint. View (c) shows the joint 90° later than (a); view (b) is the end view of the joint with angular position as in (a).

may be defined as having the same vertical height as B'_1 and b'_1 . Thus $b'_1\beta O$ and $B'_1\beta O$ are right triangles, and we may write

$$\tan \theta = \frac{b'_1\beta}{O\beta} \quad (9.5)$$

and

$$\tan \phi = \frac{B'_1\beta}{O\beta} \quad (9.6)$$

Furthermore the angle $B'_1b'_1\beta$ is a right angle, and we may write

$$\cos \alpha = \frac{b'_1\beta}{B'_1\beta} \quad (9.7)$$

Taking the ratio of Eq. (9.6) to Eq. (9.5), we have

$$\frac{\tan \phi}{\tan \theta} = \frac{B'_1 \beta}{b'_1 \beta} = \frac{1}{\cos \alpha} \quad (9.8)$$

In Fig. 9.5 is shown the variation of ϕ about its proper value as a function of θ for $\alpha = 45^\circ$. The maximum error of about 10° occurs at $\theta = 38^\circ$.

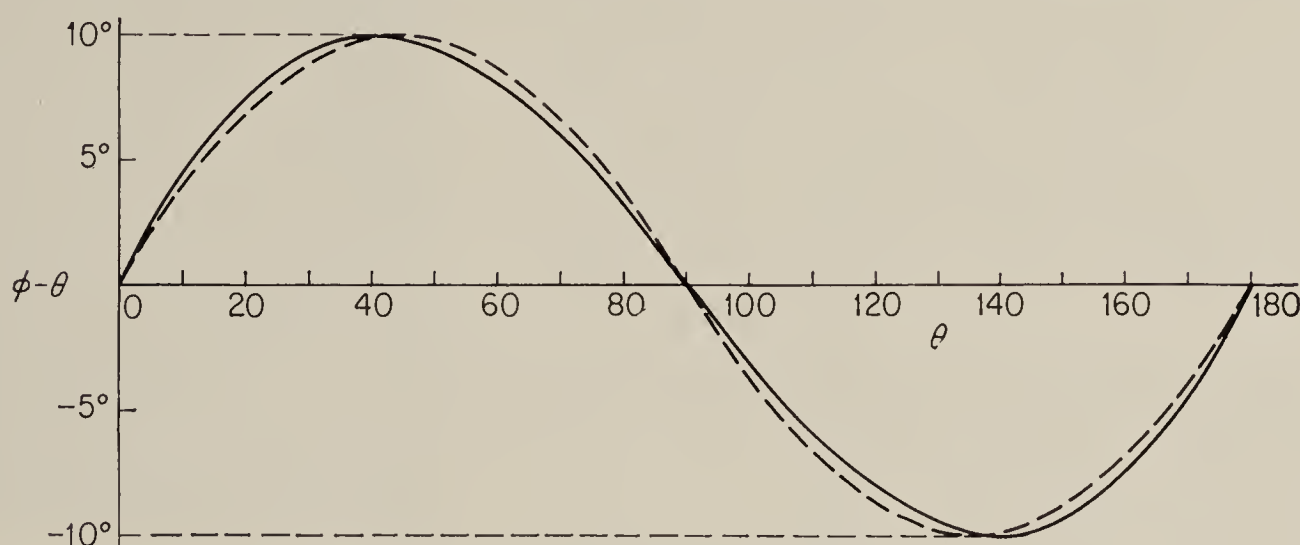


FIG. 9.5. $\phi - \theta$ versus θ for $\alpha = 45^\circ$.

In Fig. 9.6 is shown the maximum error in ϕ as a function of α and also the value of θ for which this error occurs. For small shaft angles this error is negligible. For large shaft angles two universal joints, each providing one half the total angle, should be used. If the yokes connected to the central shaft between the two universals are coplanar, the errors of the two joints can be made to cancel.

In a velocity system the angular variation in a universal joint produces spurious frequency components in the output velocity which must be considered if a careful analysis is being conducted.

9.4. Strain Gauges. A strain gauge is a device that measures minute motions of a body under applied stresses.¹ While strain is not usually of direct interest in a control system, a strain gauge may be used to measure pressure or flow in hydraulic lines, for example, by measuring the change in pipe dimensions. While completely mechanical strain gauges exist, such as extensometers, for example, strain gauges used in control applications are usually required to deliver an electric output signal. One reason

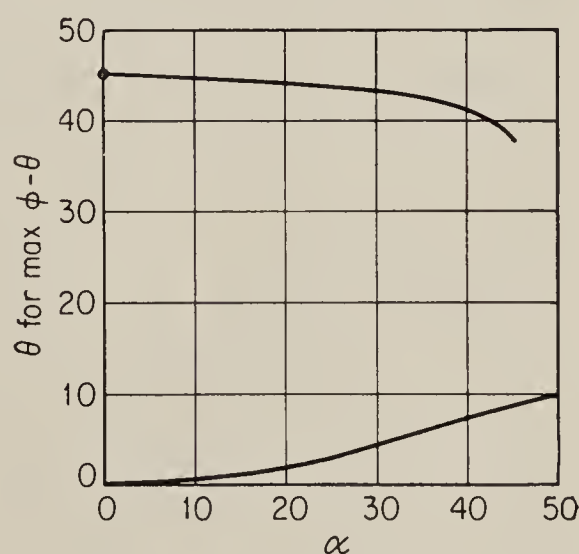


FIG. 9.6. Maximum of $\phi - \theta$ as a function of α and the value of θ for which this occurs.

¹ Hetenyi, "Handbook of Experimental Stress Analysis," John Wiley & Sons, Inc., New York, 1950.

for this is that the signal derived from a strain gauge is usually at a very low level and requires considerable amplification before it is useful. One such strain gauge is the capacitor type. It consists essentially of two plates so mounted on the object undergoing strain that they move relative

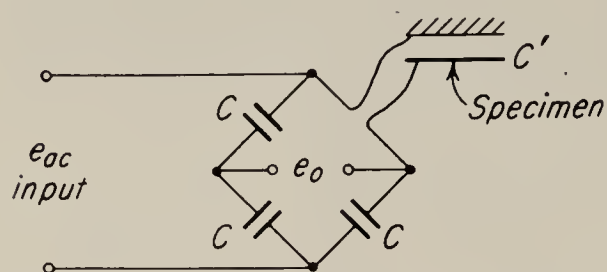


FIG. 9.7. Typical capacitor strain gauge showing bridge type of sensing circuit.

to each other as a result of the strain. The relative motion of the two plates causes a small change in the value of capacitance between them. This is best detected by the use of a Wheatstone bridge, as shown in Fig. 9.7. A constant value of a-c voltage is applied to the input terminals. A deflection of the observed body will change the spacing of the

capacitor and unbalance the bridge, thus causing a voltage to appear at the output terminals. The voltage is a function of C' and thus of strain; the relationship is given by

$$e_o = \frac{e_{in}}{2} - e_{in} \frac{C}{C + C'} = e_{in} \left(\frac{1}{2} - \frac{C}{C + C'} \right) = e_{in} \left(\frac{1}{2} - \frac{1}{1 + C'/C} \right) \quad (9.9)$$

The variation of output voltage as a function of capacitance C' is sketched in Fig. 9.8. Although the relationship between output-voltage magnitude and capacitance is nonlinear in general, the section near zero output is approximately linear. The variation of C' is usually so small that the assumption of linearity gives very accurate results. The output in Fig. 9.8 is shown positive and negative to indicate a phase reversal at null. A phase-sensitive detector is required to recover this information.

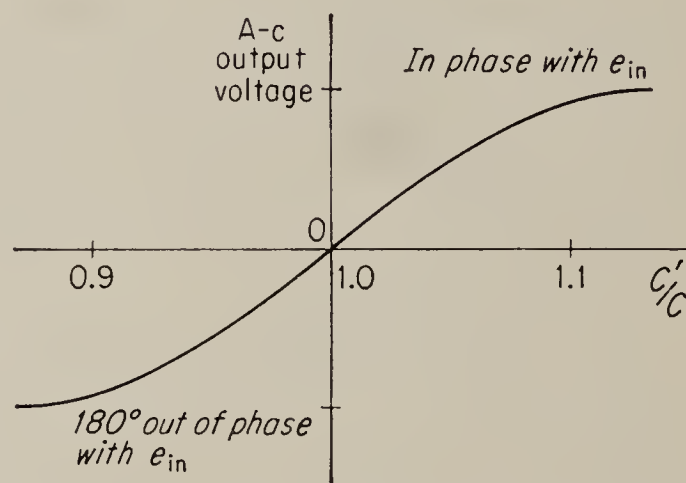


FIG. 9.8. Variation of output voltage with capacitance C' in the circuit of Fig. 9.7b.

A more common type of strain gauge is shown in Fig. 9.9a. It consists of a very fine wire, commonly nichrome, bonded to the specimen with cement and encased in plastic. As the specimen yields under stress, the nichrome wire is stretched, and its resistance increases. As in the capacitor case, the gauge is usually employed in a bridge circuit as shown in Fig. 9.9c. Although with the resistance-wire type of gauge, it is possible to use a d-c input, alternating current is usually preferred in order to simplify amplification. Whenever alternating current is used, quadrature error must be guarded against by constructing the circuit to be as symmetrical as possible with respect to ground. As in the capacitor-

bridge case, the operation may be considered linear over the normal operating range.

Symmetry can be improved, thus decreasing quadrature error, and nonlinearity decreased, as shown in Fig. 9.10, by arranging resistor *a* to

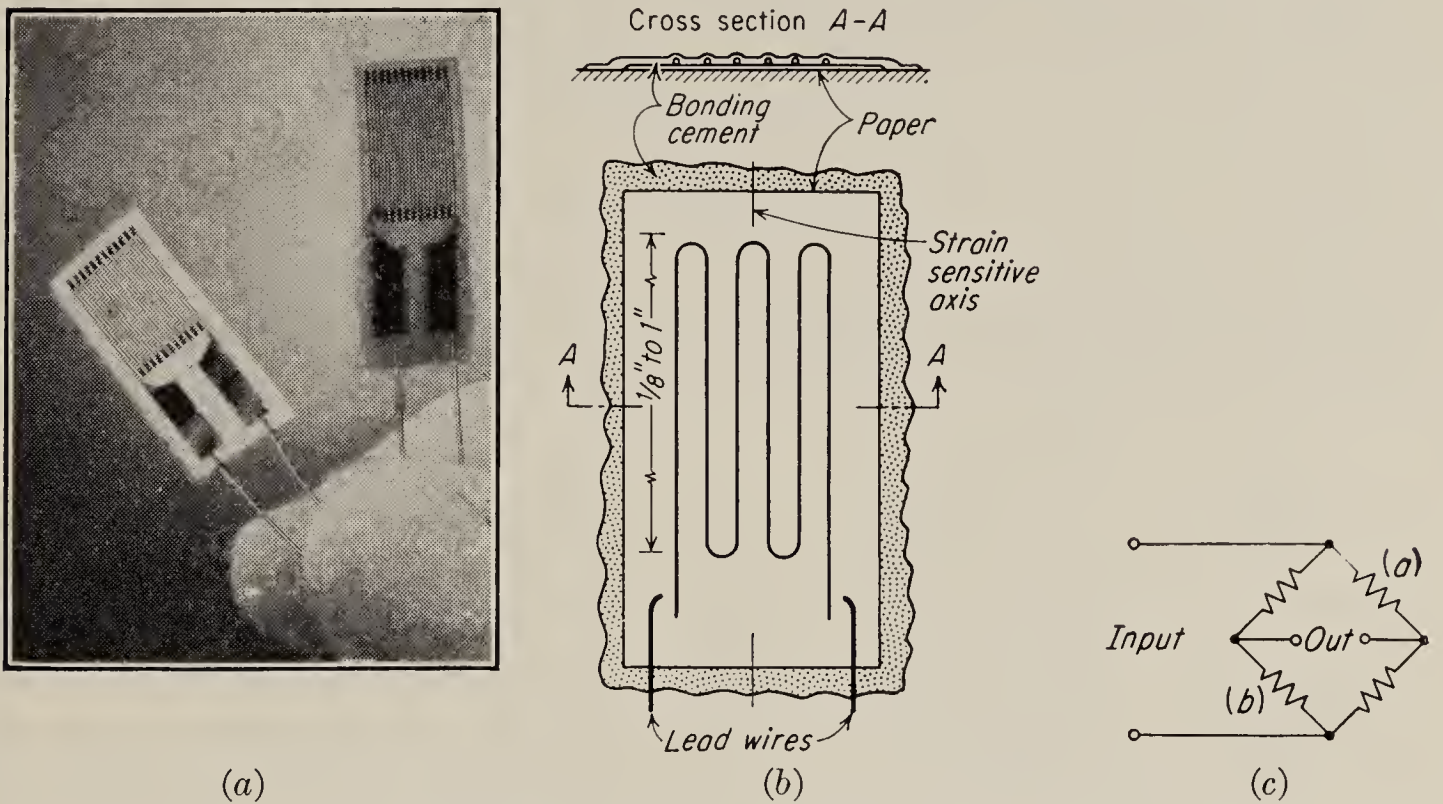


FIG. 9.9. A nichrome-wire strain gauge and bridge circuit. (Courtesy Baldwin-Lima)

stretch as resistor *b* compresses. The most serious disadvantage of nichrome-wire gauges is that they are temperature-sensitive. This sensitivity is reduced by using identical gauges in all four arms of the bridge. However, if one or two of the gauges are bonded to the specimen while the others are isolated from the specimen, temperature differentials will still occur. A possible solution is to bond all the gauges to the specimen. If two gauges are arranged to expand as the others contract, gain is increased, but the nonlinearity is also emphasized. It is sometimes possible to orient two or three of the gauges at right angles to the strain, thus immobilizing them.

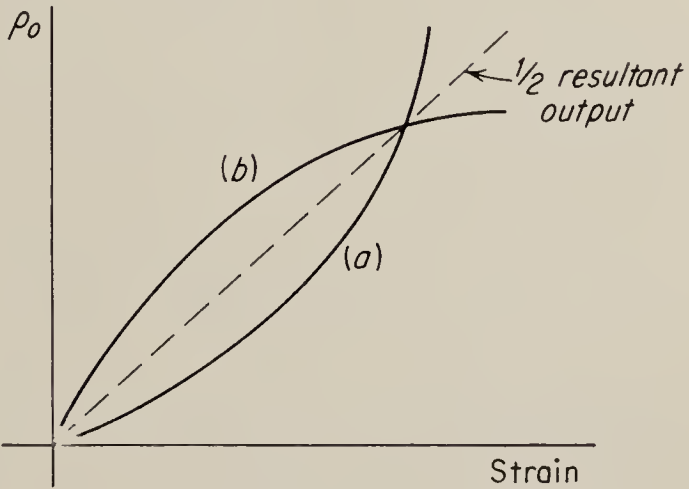


FIG. 9.10. Canceling the effect of strain-gauge nonlinearity by using two gauges.

Care must be taken that the strain gauge measures only that quantity desired. Thus, if pressure is to be measured in a pipe, vibration of the pipe must be considered because the strain gauge will also measure vibration if the vibration is allowed to reach the gauge.

9.5. The Flyweight Tachometer. One of the most reliable methods of sensing angular velocity is the flyweight tachometer. Flyweights were

the basic speed-sensing element in the flyball steam-engine governor, which is considered to be one of the first closed-loop control systems. The essential principle is still used in control systems where the requirement of absolute reliability forbids the use of electronic components.¹

Figure 9.11 shows a modern version of the flyweight tachometer. The flyweights whirl about the vertical center line and bear on the actuator, which does not turn. The spring-restrained flyweight tachometer is superior to the free tachometer because it can be made relatively free of the effects of gravity and because the weights move through a smaller arc, thus making the unit more compact. The tachometer is arranged to operate an actuator of some sort, e.g., a hydraulic valve. The position of the actuator is approximately proportional to the square of the velocity

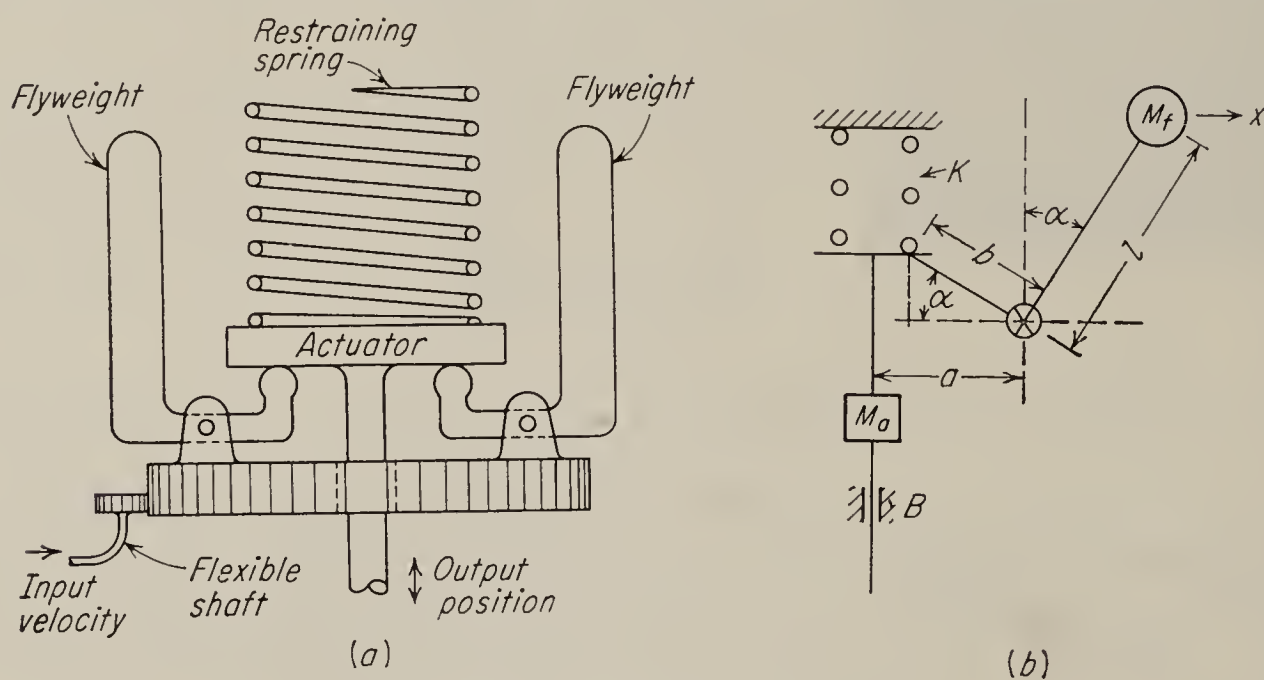


FIG. 9.11. A flyweight tachometer.

since the centrifugal force of the rotating mass increases as the square of the velocity. For the configuration shown in Fig. 9.11 the torques about the pivot point may be summed as

$$\underbrace{M_f \omega^2 (a + l \sin \alpha)}_{\text{centrifugal force}} \underbrace{l \cos \alpha}_{\text{lever arm}} = \underbrace{(Kx + B\dot{x} + M_a \ddot{x})}_{\text{reaction force}} \underbrace{b \cos \alpha}_{\text{lever arm}} \quad (9.10)$$

The various terms are defined in Fig. 9.11b:

$M_f \triangleq$ mass of flyweights

$M_a \triangleq$ mass of stem, etc.

$\omega \triangleq$ speed of rotation

$a \triangleq$ distance between pivot point and center line

$l \triangleq$ distance from pivot point to center of mass

$b \triangleq$ distance from actuator to pivot point

$K \triangleq$ spring constant

$B \triangleq$ coefficient of viscous damping

¹ For instance, in the usual hydraulic airplane-propellor pitch control.

This equation is nonlinear, not only because of the ω^2 term but also because of the presence of the $\sin \alpha$ term on the left-hand side. This term can cause nonsinusoidal oscillations under certain conditions. We shall neglect this possibility and assume that the terms in the first parenthetical expression may be represented by a “constant” K . Then taking the transform of the simplified equation,

$$\frac{\hat{x}}{\hat{\omega}^2} = \frac{M_f l K / b}{M_a s^2 + B s + K} \quad (9.11)$$

The denominator of this expression is the familiar relation for a damped spring-mass system. In practice, the system is overdamped, and the natural frequency is usually quite high compared with the response of the system in which the tachometer is installed, and hence it may be neglected. Equation (9.11) is a linear transfer function in terms of ω^2 , not ω . Thus the variation of x with ω follows a square-law relation. The restraining spring may be fashioned to counteract this effect (a so-called “hard” spring), or the operation may be assumed linear in some small operating range. Figure 9.12 shows a plot for an actual governor using a nonlinear spring. Actuator position is plotted against flyweight velocity, in the steady state. Superimposed on the curve is a square law characteristic for comparison.

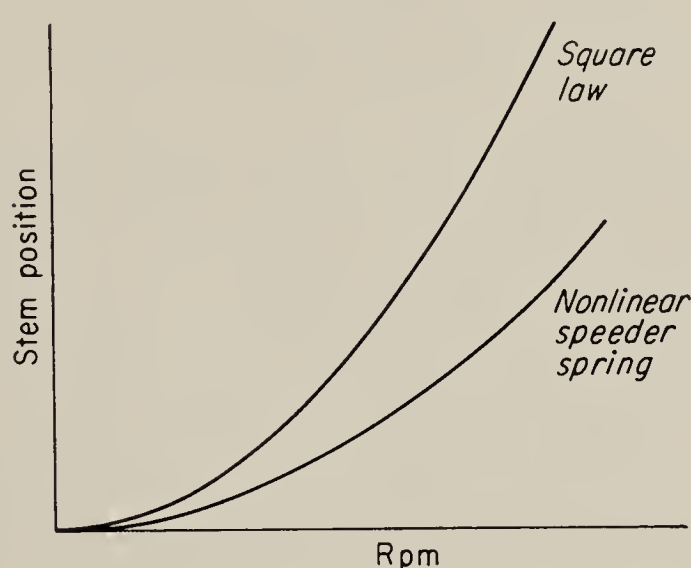


FIG. 9.12. Actuator position versus flyweight velocity for flyweight tachometer with nonlinear spring to give flatter characteristic. Square-law curve shown for comparison.

9.6. The Gyroscope. Introduction. The gyroscope is a practical embodiment of the spinning top. Basically the *gyroscope* is a disk or wheel mounted on an axle and supported by a framework, or *gimbals*, so that the disk may turn in any direction. Like the top, a gyroscope resists changes in the position of the axis about which it is spinning; an attempt to twist its axis in one direction results in a motion of the spin axis in another direction. This is called *precession* and is discussed quantitatively below.

Precession has been employed to stabilize ships and aircraft. A gyro may be mounted in such a manner that it is allowed to precess in a given plane. This precession provides a torque in a plane at right angles to the precession plane, and this torque may be used to combat the forces that cause an aircraft or ship to pitch or toss or roll. Naturally a gyro used to stabilize an ocean-going vessel or an aircraft would be quite large. For

this reason modern practice is to use the gyroscope merely as an indicating instrument or stable reference and to employ a feedback system to actuate stabilizing elements in the vehicle.

9.7. The Gyroscope. Precession. The operation of a gyroscope may be demonstrated by considering a disk spinning in free space. The disk will not change its initial angular momentum unless acted upon by outside forces. The angular velocity of the disk will be taken as ω_s . The subscript s stands for *spin axis*, i.e., the axis about which the disk is initially turning, as shown in Fig. 9.13.

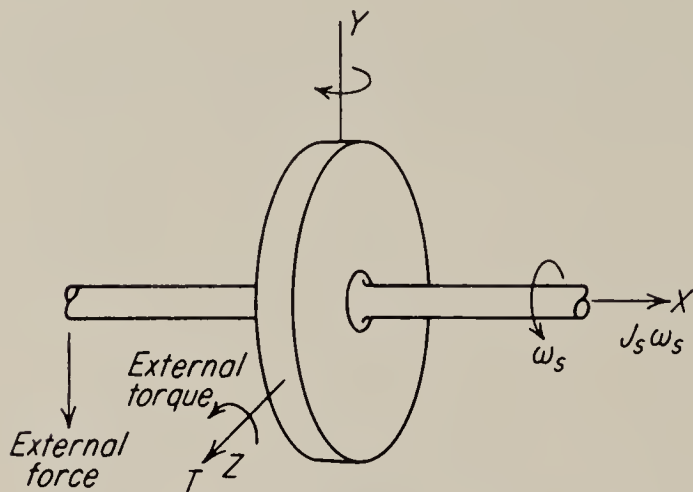


FIG. 9.13. The primitive gyroscope.

By convention, angular momentum is shown in vector form as an arrow pointing along the axis of rotation. The arrow points in the direction of advance of a right-hand screw that is turned in the same direction as the disk is turning. Now let us suppose that a torque is applied to the disk in such a direction as to attempt to turn its flat side horizontal. This torque is represented by a vector along the z axis, as shown in Fig. 9.13. By Newton's law, torque is the time rate of change of angular momentum, or

$$\mathbf{T} = \frac{d}{dt} (J\omega) = J \frac{d\omega}{dt} \Big|_{J=\text{const.}} \quad (9.12)$$

This equation is a vector equation. Since the torque is a vector pointing along the positive z axis, the change in angular velocity, $d\omega/dt$, must also point in the positive z direction, since J is not a vector quantity. This is shown in the vector diagram of Fig.

9.14, in which the y axis is taken to be perpendicular to the plane of the paper. We see that the effect of the torque T_z is to rotate the spin axis of the gyroscope about the y axis. This rotation is the gyroscopic precession.

Note that torque applied along the positive z axis causes the spin axis of the gyroscope to rotate in the negative y direction. Thus the spin axis tends to line itself up with the torque vector.

Figure 9.14 can be used to derive the quantitative relation between the torque and the resulting velocity of precession. Thus let the torque T_z act for a short increment of time, dt . Its effect is to rotate the spin axis through the small angle $-d\theta_y$. If the angle is very small, it is equal to its

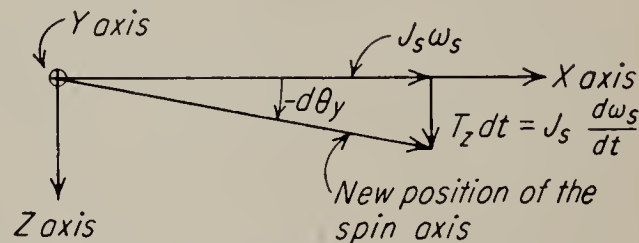


FIG. 9.14. Gyroscope vector diagram.

tangent, or

$$-d\theta_y = \tan(-d\theta_y) = \frac{T_z dt}{J_s \omega_s} \quad (9.13)$$

so that

$$\frac{d\theta_y}{dt} = \omega_y = -\frac{T_z}{J_s \omega_s} \quad (9.14)$$

where ω_y is the velocity of precession.

By similar reasoning it can be shown that the gyro precesses about the z axis when a torque is applied along the y axis, or

$$\frac{d\theta_z}{dt} = \frac{T_y}{J_s \omega_s} \quad (9.15)$$

Again the direction of precession is such as to line up the spin axis with the direction of the torque vector.

Equations (9.14) and (9.15) are only approximations, because they do not consider the fact that a change in precession velocity, multiplied by the rest inertia of the gyro, also represents a change in angular momentum, which requires torque. Also they neglect the effects of gimbal friction, which result in braking torques when the precession velocities become large. The equations are therefore reasonably accurate only as long as the precession velocity is low and the torque variations small.

There are several interesting things about the results obtained thus far. First, the precession is along an axis perpendicular to that of the applied torque; yet had not the disk a velocity about the spin axis, rotation would have taken place along the axis of the applied torque. Second, the applied torque results in a velocity rather than an acceleration, and finally the moment of inertia about the spin axis enters into the relation rather than the moment of inertia about the axis of the applied torque. All of these facts make the gyroscope somewhat puzzling on first examination.

The phenomenon of precession may be examined in another manner that is perhaps more closely related to familiar concepts. To show the force required for precession, the velocity of precession will be assumed, and the required force will be found. The coordinate system fixed in space is as shown in Fig. 9.15. Assume that the disk is spinning with an angular velocity ω_s and also that the disk is precessing with a velocity ω_p . Consider a particle of the disk near the rim.¹ This particle has a linear velocity parallel to the x axis, which is due only to the precession. This linear velocity of the particle varies as the particle is carried around by the spin velocity. At point A in space the component of velocity in the x direction has a maximum negative magnitude, and at point C the

¹ S. T. Preston, The Mechanics of the Gyroscope, *Sci. American Supplement*, vol. 58, no. 1501, pp. 24,057–24,058, Oct. 8, 1904.

velocity has a maximum positive magnitude. At points B and D the linear velocity is zero. Figure 9.15*b* shows a pictorial representation of the linear velocity in the x direction of a particle of the disk. If we now consider the linear acceleration of the particle in the x direction, we find that at point B the velocity is changing from negative to positive and the acceleration has a maximum positive magnitude. Throughout the semi-circle ABC the acceleration is positive, and on the upper half of the disk,

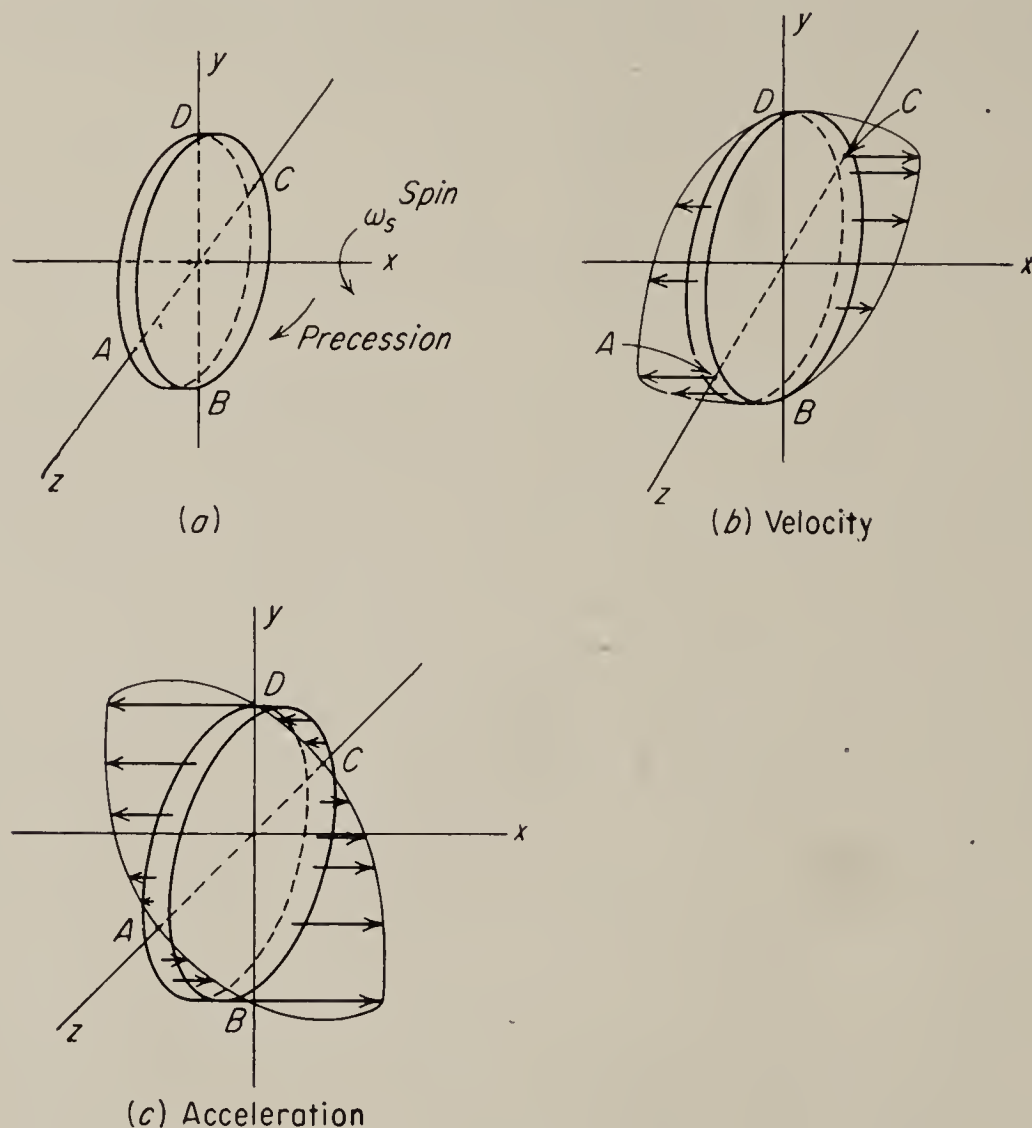


FIG. 9.15. (a) The disk with respect to coordinates fixed in space. (b) The linear velocity of a particle of the disk in the x direction. (c) The linear acceleration in the x direction of a particle of the disk.

the acceleration is negative, as shown in Fig. 9.15*c*. By Newton's second law a force must be applied to the disk in order to cause this acceleration, or more conveniently it may be applied to the axle, as was shown in Fig. 9.13.

This explanation of precession has the advantage of utilizing only Newton's second law in its simplest form. Furthermore the acceleration is in the direction of the force, which appeals to reason, and finally the applied force results in an acceleration.

It is possible to sum up the forces on all the particles to obtain the total torque required to produce a given velocity of precession. Den Hartog¹

¹ Den Hartog, "Mechanics," McGraw-Hill Book Company, Inc., New York, 1948.

has shown that this process gives results identical to those already obtained.

9.8. Gyroscope. Equations of Motion. When the gyroscope is used as a control-system component, the elementary relations between torque and precession velocity [Eqs. (9.14) and (9.15)] are not sufficiently accurate, and the effects of gyro rest inertia and gimbal friction must be considered. This may be done rather easily if the arrangement of the gyro disk, gimbals, and torque motors is such that during precession the spin axis always remains in a plane perpendicular to the torque causing the precession. A commonly used configuration meeting this requirement is shown in Fig. 9.16. Note that the torque of one of the torque motors

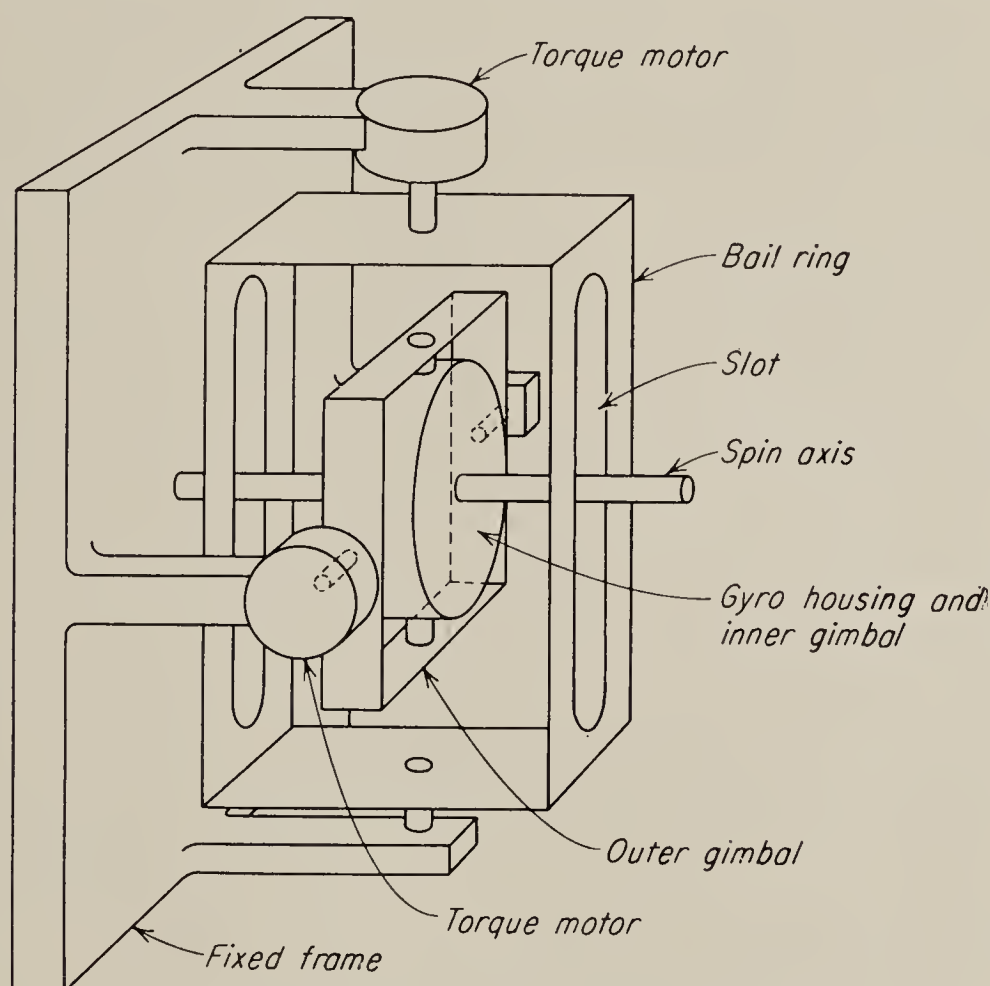


FIG. 9.16. Gyro with bail-ring type of gimbal.

is applied to the gyro through a *bail-ring* having a slot in which the gyro spin axis is free to slide. This arrangement permits both torque-motor stators to be fixed to the outer frame. In this type of system we may consider the torques required to accelerate the gyro rest inertia and to overcome gimbal friction as subtracting from the “applied” torque, leaving an “effective” torque resulting in precession only. If the gimbal friction is assumed to be viscous, then this reasoning modifies Eqs. (9.14) and (9.15) as follows:

$$T_z - J_z \frac{d^2\theta_z}{dt^2} - B_z \frac{d\theta_z}{dt} = -J_s \omega_s \frac{d\theta_y}{dt} \quad (9.16)$$

$$T_y - J_y \frac{d^2\theta_y}{dt^2} - B_y \frac{d\theta_y}{dt} = J_s \omega_s \frac{d\theta_z}{dt} \quad (9.17)$$

where J_z and J_y are the rest inertias of the gyro and gimbals measured in the z and y axes, and B_z and B_y are the coefficients of friction in these axes. The torques T_z and T_y are the components of torque perpendicular to the spin axis, since components parallel to the spin axis do not result in precession. Hence if the actual torques applied by the torque motors are T'_z and T'_y and if θ_z and θ_y are measured between the spin axis and a line perpendicular to the two torque axes, then Eqs. (9.16) and (9.17) become

$$T'_z \cos \theta_y - J_z \frac{d^2 \theta_z}{dt^2} - B_z \frac{d\theta_z}{dt} = -J_s \omega_s \frac{d\theta_y}{dt} \quad (9.18)$$

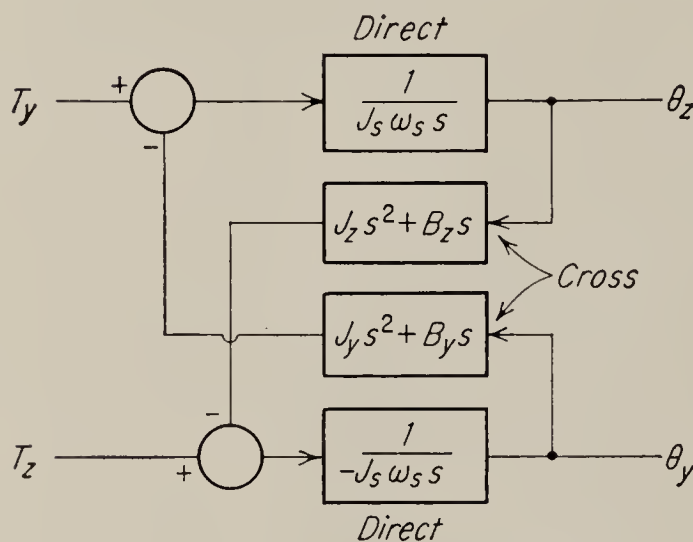
$$T'_y \cos \theta_z - J_y \frac{d^2 \theta_y}{dt^2} - B_y \frac{d\theta_y}{dt} = J_s \omega_s \frac{d\theta_z}{dt} \quad (9.19)$$

The trigonometric functions multiplying the applied torque clearly make this a nonlinear set of differential equations which is difficult to solve in

general. In particular, it is not possible to obtain transfer functions relating the torque applied by a torque motor to displacement of the gyro spin axis.

In many cases of practical interest, it is true, however, that θ_y and θ_z are approximately zero; i.e., the gimbals of the gyro of Fig. 9.16 are at right angles to each other and to the frame. In that case the cosine functions appearing in Eqs. (9.18) and (9.19) are approximately equal to unity, and the equations become linear. They can then be transformed

FIG. 9.17. An approximate block-diagram representation of a gyroscope.



and rewritten as follows:

$$(J_z s^2 + B_z s) \hat{\theta}_z - J_s \omega_s s \hat{\theta}_y = \hat{T}_z \quad (9.20)$$

$$(J_y s^2 + B_y s) \hat{\theta}_y + J_s \omega_s s \hat{\theta}_z = \hat{T}_y \quad (9.21)$$

A block diagram may be derived from these equations; one form is shown in Fig. 9.17.

When gimbal arrangements differing from that shown in Fig. 9.16 are used, the behavior of the gyroscope may become more complicated. Thus, consider the arrangement shown in Fig. 9.18. Here the stator of the torque motor driving the inner gimbal is mounted on the outer gimbal. Careful inspection of this figure will indicate that the gyro may precess in such a way that the spin axis traces out a conical surface rather than a plane. This happens, in fact, whenever torque is applied to the inner gimbal by the torque motor mounted on the outer gimbal, and when the two gimbals are not at right angles to each other.

The differential equations describing the motion of this type of gyroscope cannot be obtained by the simple reasoning employed in obtaining Eqs. (9.18) and (9.19). They are nonlinear equations and contain terms involving the product of the two precession angles in addition to trigonometric terms. For a detailed discussion of gyroscopes operating in this way, the reader is referred to texts on advanced dynamics.¹ Here again, if the assumption is made that the gimbals are usually approximately perpendicular to each other and to the frame, no difficulty is experienced in linearizing the differential equations; in fact they become identical with Eqs. (9.20) and (9.21) derived above.

The transfer function relating precession angle to applied torque can be obtained by inspection of the block diagram of Fig. 9.17, or by a simultaneous solution of Eqs. (9.20) and (9.21):

$$\frac{\hat{\theta}_z}{\hat{T}'_y} = \frac{J_s \omega_s}{s[J_y J_z s^2 + (J_y B_z + J_z B_y)s + (J_s \omega_s)^2 + B_y B_z]} \quad (9.22)$$

This transfer function is, of course, accurate only if the conditions under which Eqs. (9.20) and (9.21) were derived are met. Normally, friction is sufficiently small to make $B_y B_z \ll J_s \omega_s$; hence Eq. (9.22) is approximately equivalent to

$$\frac{s\hat{\theta}_z}{\hat{T}'_y} = \frac{1/J_s \omega_s}{\frac{J_y J_z}{J_s^2 \omega_s^2} s^2 + \frac{J_y B_z + J_z B_y}{J_s^2 \omega_s^2} s + 1} \quad (9.23)$$

We note that the gyro response has the standard quadratic form, and if friction is small, the response to step or other shock inputs is oscillatory. The transient oscillations observed in a gyro are referred to as *nutations*, since the gyro spin axis characteristically traces out a conical surface during these oscillations. The *frequency of nutation*, i.e., the undamped natural frequency of Eq. (9.23), is seen to be

$$\omega_n = \omega_s \frac{J_s}{\sqrt{J_y J_z}} \quad (9.24)$$

¹ See, for instance, Page, "Introduction to Theoretical Physics," D. Van Nostrand Company, Inc., Princeton, N.J., 1935, pp. 137-147.

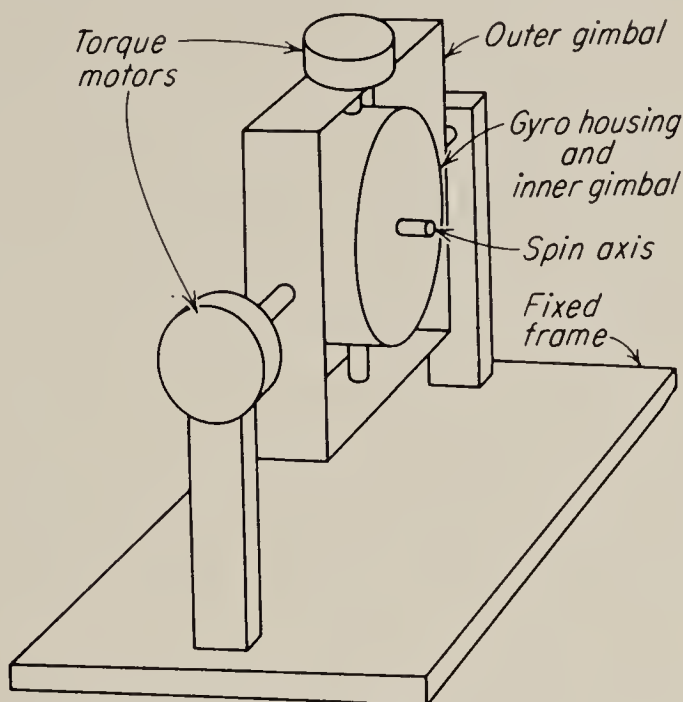


FIG. 9.18. Gyro capable of precessing along a conical surface.

Since J_y and J_z are usually larger than J_s , the nutating frequency is usually less than the spin velocity. In fact it can be demonstrated that the nutation frequency of a thin disk spinning in space without gimbals is equal to one-half the spin velocity.

The transfer function relating precession velocity to precessing torque [Eq. (9.23)] is referred to as the *direct transfer function* of the gyro. However, owing to the gyro rest inertia and gimbal friction, a torque applied to one set of axes of a gyro also results in some gyro motion in the same axis. The relation between torque and angular displacement along the same axis is referred to as the *cross-coupling transfer function* and can be

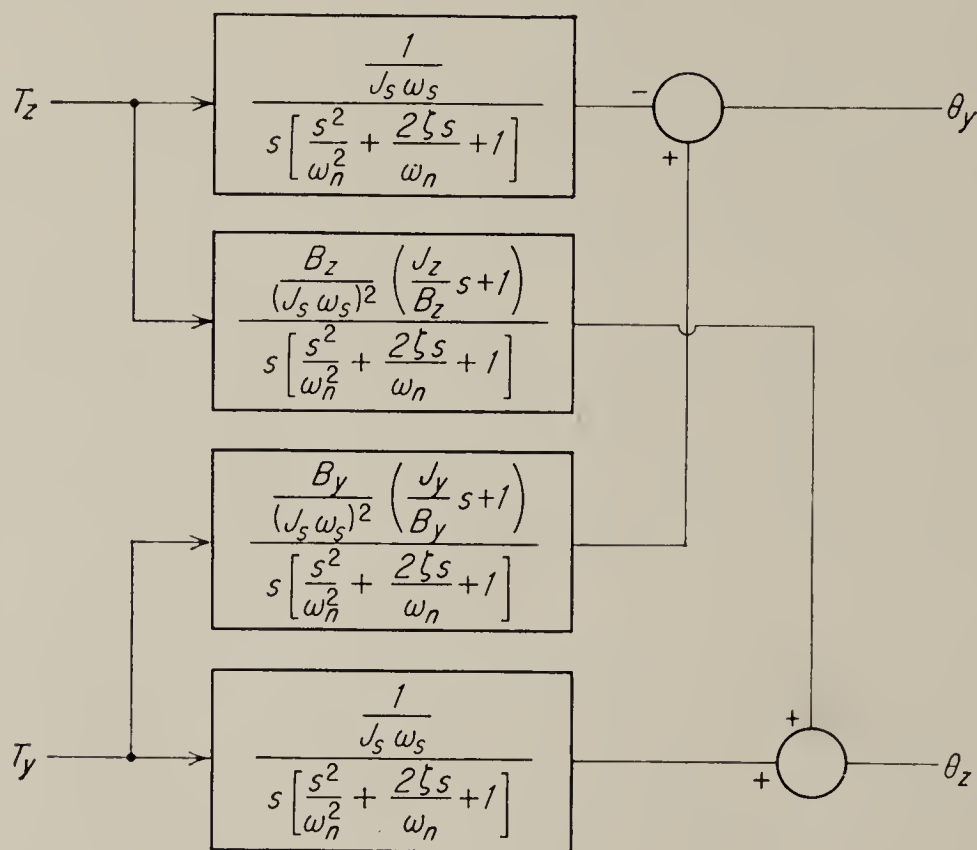


FIG. 9.19. A more complete gyro block diagram.

derived from Eqs. (9.20) and (9.21). If gimbal friction is small, it is given approximately by

$$\frac{\hat{\theta}_z}{\hat{T}_z} = \frac{B_y[(J_y/B_y)s + 1]}{J_s^2 \omega_s^2 s \left(\frac{J_z J_y}{J_s^2 \omega_s^2} s^2 + \frac{J_z B_y + J_y B_z}{J_s^2 \omega_s^2} s + 1 \right)} \quad (9.25)$$

A similar expression may be derived for the y axis. A large cross coupling is usually undesirable in a gyro, and Eq. (9.25) indicates that it is minimized by reducing friction and by making the spin velocity as large as possible.

The complete set of gyro transfer functions may be summarized in the block diagram shown in Fig. 9.19. In this figure ω_n is the nutation frequency given by Eq. (9.24), and ζ is the damping coefficient, which, by Eq. (9.23), is equal to $(1/2J_s \omega_s)(B_z \sqrt{J_y/J_z} + B_y \sqrt{J_z/J_y})$.

9.9. Practical Gyroscopes. The rotor of a practical gyroscope is not usually a disk but instead consists of the rotor of a polyphase induction motor designed to operate at speeds as high as 20,000 rpm. Some of the early aircraft instruments were, however, disks driven by a jet of air. Some of the modern aircraft gyros are enclosed in a temperature-stabilized environment and may be hermetically sealed. The sealed case may be supported or floated in a damping fluid if demanded by the application (see Sec. 9.11). Usually there is provision made in aircraft gyros to lock or *cage* the gyro so that it cannot precess or move in any direction. This is necessary so that, when the craft undertakes violent maneuvers, the gyro will not be driven into its stops or wound up in its flexible connections. The present manufacturing trend in aircraft gyros is to employ a separate gyro in a single gimbal for each coordinate to be sensed. This eases the problems of erection, or placing the gyro spin axis at its desired setting, and the problem of providing precession torque and of sensing gyro position.

A major problem in the construction of practical gyros is the minimization of drift. Drift is caused by a number of factors, but the primary causes are unbalance of the gimbals and bearing friction. Unbalance of the gimbals results in gravitational forces, which tend to precess the gyro. The vibration that is always present when the gyro is spinning acts to nullify the small static friction present even in the finest bearings, so that mass unbalances which would not be detectable when the gyro disk is not running will produce small rates of precession when the gyro is spinning. Drift from this source can be minimized by fitting the gimbals with adjustable weights so that any slight unbalance may be compensated for.

Bearing friction results in precessing torques whenever the external mounting of the gyro is shifted. Bearing friction is caused by small irregularities in the bearing surfaces and is largely random in character. Hence it results in random precession, or drift, which usually has a non-zero average value. The average amount of drift varies with the installation. With small gyros such as those used in artificial horizons of light airplanes, the drift is often large enough to require manual resetting of the gyro every 10 minutes. On the other hand, very high quality gyroscopes can be built with average drift rates of less than 1° in 12 hours. Prior to World War II, such gyros required very heavy rotors, a typical model designed for shipboard installation having a rotor weighing 54 lb. Present manufacturing procedures have, however, developed to the point where a single-degree-of-freedom gyro, weighing about 5 lb complete with torque motors and case, can now be obtained with approximately the same drift.

Probably the most critical example of gyroscope applications occurs in inertial navigation systems. In these systems the gyroscope is usually

a primary reference although systems that are corrected by optical star-tracking methods have been proposed. In any case the simpler magnetic and gravity references cannot be used. A magnetic reference is useless in transpolar flights, and gravity is undesirable as a reference since it is not constant around the globe. Obviously both gravity and terrestrial magnetism would be useless as references in interplanetary flight.

To minimize drift, gyros must be constructed with extremely fine bearings and have, therefore, almost no inherent damping. For this reason, when a free gyro is used as part of a feedback loop, as in the gyroscopic velocimeter described in Sec. 9.10, it contributes roots to the characteristic equation of the system that are almost on the imaginary axis. Some form of equalization is therefore always required in such feedback circuits to produce a stable and well-damped system. In rate gyros, the lack of inherent damping results in a highly oscillatory response, as indicated by Eq. (9.27). For this reason some form of external damping must usually be added. In some models electromagnetic damping is used; others are damped by enclosing the gyro in a damping fluid; this latter scheme simultaneously provides a shock mounting for the device. It should be noted that purely viscous damping should theoretically not result in any average drift, since it is nonrandom and perfectly symmetrical. Furthermore, damping devices may be constructed (somewhat on the order of the Lancaster damper described in Sec. 8.5) which dissipate energy whenever the member to which they are attached moves, but which add no friction between the gyro and the stationary frame.

Proper damping is particularly important in marine gyrocompasses. In addition to being built with the finest bearings available, these gyros are usually constructed to have an undamped natural period of the order of 85 minutes. This long period, which incidentally is the period of a simple pendulum having a length equal to the radius of the earth, has been shown by Schuler¹ to result in a compass whose north-south indication is not affected by accelerations of the ship or airplane on which it is mounted. A gyro having this period is said to be *Schuler tuned*, and it is clear that any device having such a long period should be damped so as to cause the oscillations to die out as quickly as possible. The damping problem is, however, complicated by the fact that some of the more common damping methods may cause north-seeking errors,² and a number of special methods are therefore used. For detailed description of these methods the reader is referred to textbooks dealing with gyrocompasses.³

¹ M. Schuler, Die Störung von Pendel und Kreisel apparaten durch die Beschleunigung des Fahrzeuges, *Physik. Z.*, vol. 24, p. 344, 1923.

² Rawlings, "The Theory of the Gyroscopic Compass," The Macmillan Company, New York, 1944.

³ Richardson, "The Gyroscope Applied," Philosophical Library, Inc., New York, 1954.

9.10. Gyroscope Applications. Two basic gyro types are used in practice: the free gyro and the restrained gyro. A free gyro is mounted in a set of gimbals so that angular motion of the supporting frame is not transmitted to it. Ideally the direction of the spin axis of a free gyro is therefore fixed in space. Such gyros are used to provide a fixed reference position, for instance, in the vertical and directional gyros found in automatic pilots and in inertial navigation equipment. Other uses of this sort are in artificial horizons, in roll-stabilizing equipment, and so forth.

Since even the most carefully constructed gyros are subject to a certain amount of random drift, the accuracy of the reference direction established by the gyro tends to deteriorate with time. In some cases this is of no consequence. Thus, a gyro used to establish the vertical or directional references in a guided missile having a life of only a few minutes may drift a few degrees per hour without any adverse effect on the accuracy of the missile. Drift in the gyro used as the artificial horizon in an airplane may not be important if the gyro can be reset occasionally. However, in more critical applications the gyro is not completely free but is equipped with small torque motors so that it can be precessed. In this way the gyro can be automatically slaved to a more accurate primary reference. Thus, a gyro used to establish a vertical may be slaved to a pendulum of some sort. The position of the gyro is continuously compared to that of the pendulum, and any difference in the positions is used to precess the gyro to make its position correspond to that of the pendulum. The precessing torque applied to the gyro is kept so small that the maximum precession rate is only a little larger than the maximum drift expected. In this way the gyro cannot follow rapid fluctuations of the pendulum position, and the system acts merely to keep the gyro spin axis aligned with the average pendulum position. Thus the gyro acts not so much as a primary reference but as an integrating or smoothing device applied to the actual reference. In a similar way directional gyros are often slaved to a magnetic compass which provides the primary north-south reference.

A somewhat different principle is used in the gyrocompass used on ships. This is essentially a free gyro equipped with a weight that tends to keep the spin axis in a horizontal plane. The weight hangs, supported by two bearings on the spin axle, centered below the gyro disk. This weight, together with the rotation of the earth, acts to precess the gyro until the spin axis takes on a north-south position.

An application of the gyroscope that differs fundamentally from those described above is the measurement of low angular velocity. In this application, use is made of the fact that the precession velocity is directly proportional to the precessing torque. Thus in one form of gyroscopic velocimeter the gyroscope is equipped with synchro devices (see Chap. 5)

that sense the relative displacement between the gimbals and the outer frame. The output of the synchro is amplified and applied to the proper torque motor in such a way that the resulting gyro precession returns the output of the synchro to zero. This feedback system therefore acts to

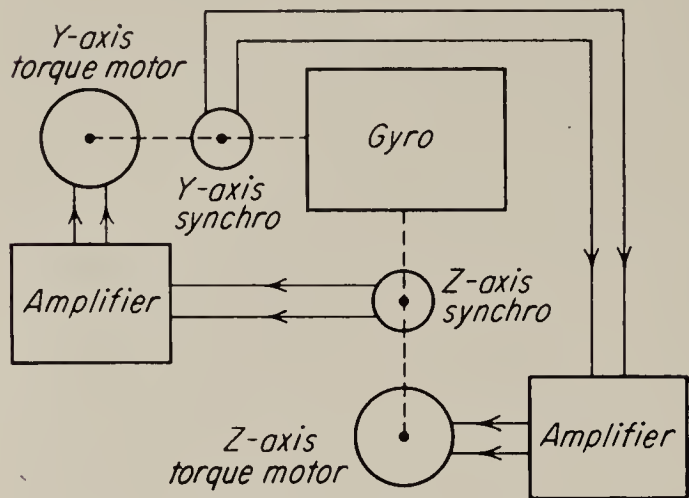


FIG. 9.20. Gyroscopic velocimeter.

keep the gyroscope aligned with the outer frame. The torque developed by the torque motors to accomplish this result is, however, by Eq. (9.14) directly proportional to the precession velocity and therefore to the velocity of the outer frame. Commonly, the torque motor is an electric motor, and the torque is proportional to the voltage applied across the windings (see Chap. 7). This voltage may there-

fore be used as an indication of the velocity. A simplified schematic diagram of such a system arranged to measure velocity in two axes is shown in Fig. 9.20.

A similar, although somewhat simpler, device working on essentially the same principle is the *restrained*, or *rate gyro*, shown in Fig. 9.21. In

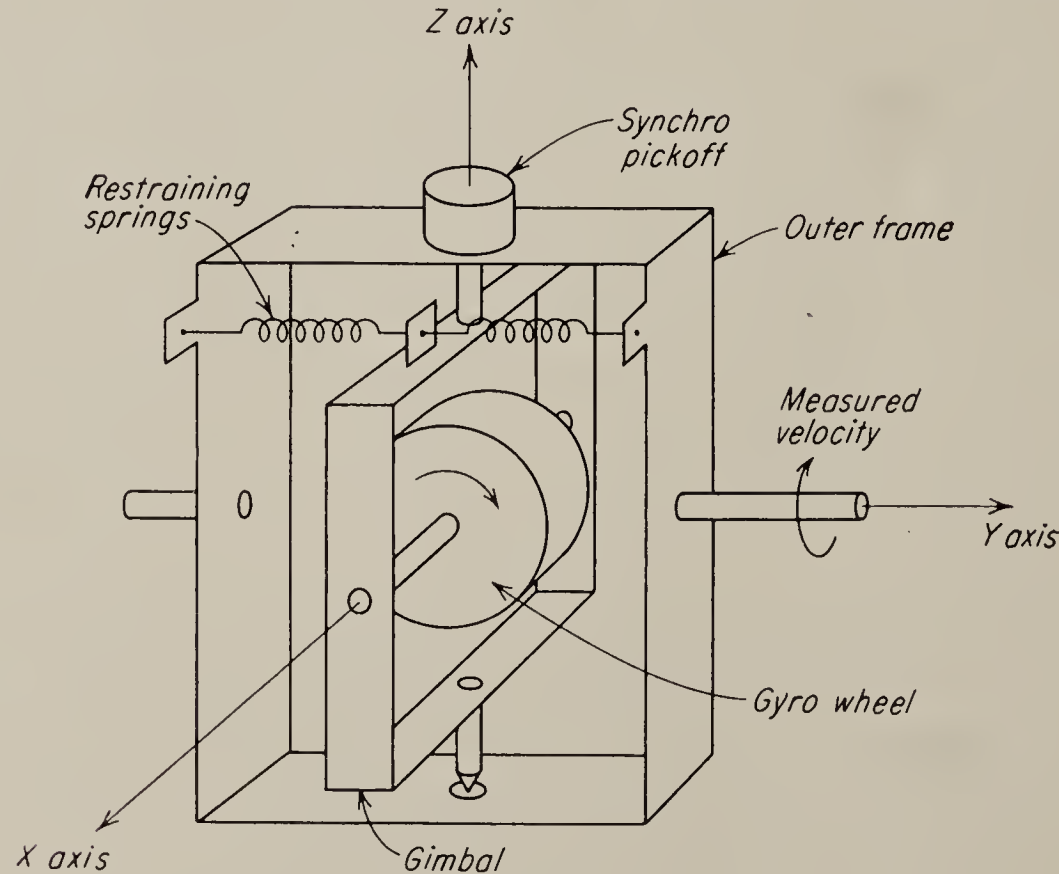


FIG. 9.21. A simplified schematic of a rate gyro.

this figure the gyro wheel is mounted in a gimbal which is pivoted so as to be able to rotate in an outer frame. The rotation of the gimbal relative to the outer frame is, however, checked by means of the restraining

springs shown. Suppose now that the entire assembly is rotated about the y axis. Y -axis torque is applied to the gyro through the gimbal, and the gyro begins to precess about the z axis. This precession is arrested by the springs, which act to produce a torque on the gyro about the z axis, opposing the z -axis precession. The torque generated by the springs results in precession about the y axis, so that the gyro is able to follow the rotation of the outer frame. The torque supplied by the springs must, therefore, be precisely sufficient to precess the gyro at the rate at which the outer frame is being rotated. Hence this torque is a direct measure of the speed of the outer frame. The simplest way to measure the torque is to measure the change in displacement of the gimbal relative to the outer frame, since if the restraining springs are linear, this displacement is proportional to the torque. For maximum accuracy the spring should not deflect appreciably, since the torque required to precess the gyro about the y axis (fixed with respect to the outer frame) becomes smaller as the spin axis of the gyro approaches the y axis.

If the angular momentum of the gyro is large, a large torque is generated by relatively small velocities. Hence rate gyros can be made sensitive to very low angular velocities. With reasonable care a rate gyro can be constructed that indicates the angular velocity of the earth, a velocity of less than 0.0007 rpm. It should be noted, incidentally, that the velocity measured by a rate gyro is always with respect to an inertial reference fixed in space; furthermore since velocities about the x and z axes of Fig. 9.21 do not have the action described, the rate gyro is in general sensitive only to the y component of the angular velocity applied to it.

The transfer function of the rate gyro can be obtained from Eq. (9.20). The spring acts to make T_z proportional to θ_z ; hence Eq. (9.20) may be rewritten as follows:

$$K_z \hat{\theta}_z + B_z s \hat{\theta}_z + J_z s^2 \hat{\theta}_z = J_s \omega_s s \hat{\theta}_y \quad (9.26)$$

where K_z is the spring constant of the restraining spring. Hence

$$\hat{\theta}_z = \frac{J_s \omega_s}{J_z s^2 + B_z s + K_z} s \hat{\theta}_y \quad (9.27)$$

Note that some damping is necessary in the z axis to prevent an oscillatory response.

9.11. Application of Gyroscopes to Inertial Navigation. Inertial navigation is a rather recent development which requires gyroscopes with accuracies that would have been undreamed of a few years ago. Although gyros of many types are used in this application, one of the most frequently used types is a single-degree-of-freedom type developed by C. S. Draper and associates. This is the so-called HIG gyro (hermetic

integrating gyro).¹ A schematic diagram of this gyro is shown in Fig. 9.22. In principle the construction of this gyro is quite similar to that of the simple rate gyro shown in Fig. 9.21; however, no restraining springs are used. Also, the inner gimbal, which holds the spinning gyro wheel, is built in the form of a hermetically sealed cylindrical can, or float, with shaft extensions. The can is filled with helium which acts as a neutral atmosphere. The outer case of the gyro is completely filled with a viscous fluid whose density is just sufficient so that the gimbal and gyro wheel are in neutral buoyancy; i.e., in a large quantity of the fluid the inner can would neither sink nor rise to the surface. Thus the load on the jewelled bearings that support the gimbal to the outer case is negligibly small, and there is therefore practically no friction between the outer case and the gimbal. Attached to one of the shafts extending from the gimbal can is the rotor of a microsyn generator (see Sec. 5.9), and a torque motor is attached to the other shaft.

The operation of this gyro may be explained as follows. Suppose that the outer case is rotated about the y axis (see Fig. 9.22). As has been explained previously, the gyro will then tend to precess about the z axis. There are no springs to arrest the precession as in the rate gyro, but since the inner gimbal is floated in a highly viscous fluid, a viscous shear torque is developed which limits the precession velocity. Thus in Eq. (9.27) the spring constant term vanishes, and we obtain for the transfer function:

$$\hat{\theta}_z = \frac{J_s \omega_s / B_z}{(J_z / B_z)s + 1} \hat{\theta}_y$$

The output angle θ_z is detected by the microsyn generator mentioned above, and since in most practical applications the variation of the input angle is quite slow the time constant J_z / B_z is usually sufficiently small as to be negligible. Hence for small input angles θ_y the output signal is proportional to the input angle rather than to the rate of change of input angle. This is the reason for the name “integrating gyro.”

Note that the HIG gyro furnishes the same sort of information as a free gyro, i.e., attitude information. However, the special construction features, in particular the fact that it is a single-axis device, make this gyro much more accurate and rugged than the usual free gyros.

It should be noted that although the primary input axis to the gyro is the y axis of Fig. 9.22, the gyro will also deliver an output signal when the case is accelerated about the z axis. This type of output is usually undesirable and is minimized by using a highly viscous hydraulic fluid. The

¹ C. S. Draper, W. Wrigley, and L. R. Grohe, The Floating Integrating Gyro and Its Application to Geometrical Stabilization Problems on Moving Bases, *Aeronaut. Eng. Rev.*, vol 15, no. 6, June, 1956.

developers of the device state that for the usual accelerations of the z axis encountered in practice, this output is negligible.

In most applications to inertial navigation three HIG gyros are mounted on a stable platform to detect rotation about the x , y , and z axes. The output signal from each gyro is amplified and applied in the proper fashion to servomotors which control the rotation of the stable platform about the three axes. Thus, the platform is maintained at a fixed attitude with respect to inertial space.

The torque motors mounted on the shaft extension of the inner gimbal of the HIG gyro may be used to apply an additional precessing torque to the gyro. This torque will produce an output signal even when the input

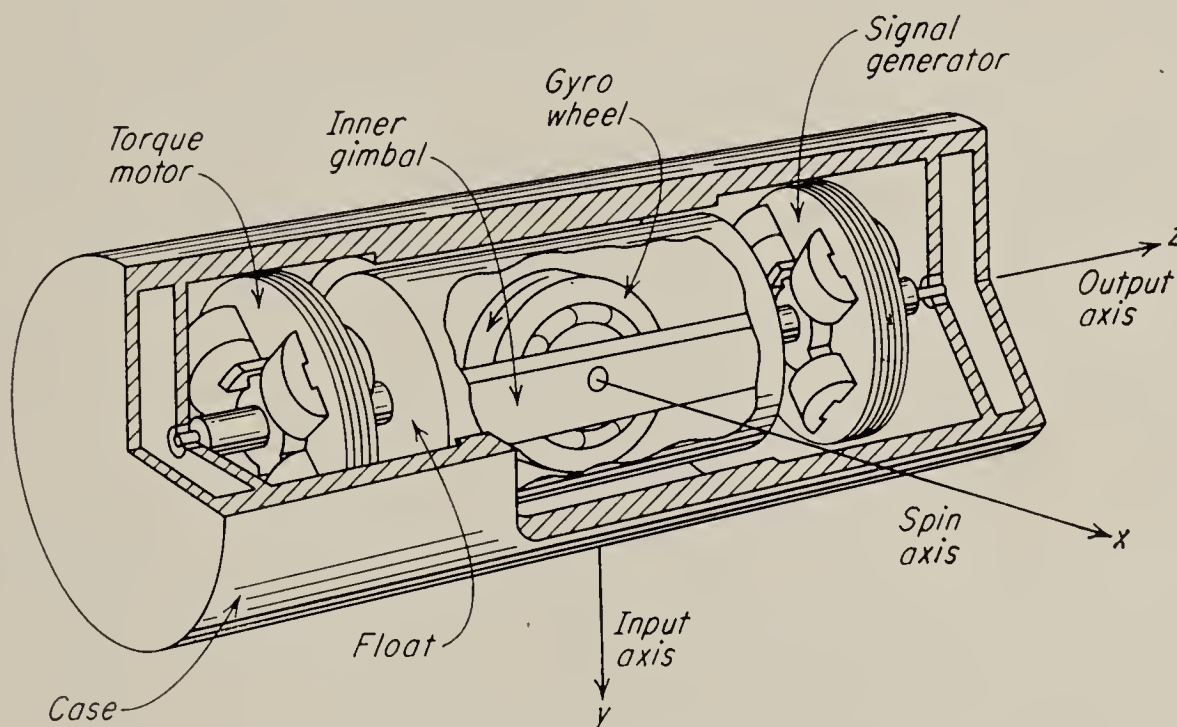


FIG. 9.22. Simplified diagram of the HIG gyro.

angle θ_y is zero. This output will be amplified and applied to the servomotor with the result that the stable platform tilts, or turns, at a rate proportional to the torque input. Thus the platform position can be controlled by currents applied to the torque motor. It is necessary to do this since the gyros will tend to maintain the platform fixed with respect to an inertial reference in space. It is, however, normally desirable that the platform be maintained at a fixed attitude (usually horizontal) with respect to the earth. Since the earth rotates in space a platform mounted on a fixed point on the earth's surface would therefore have to be rotated at the same rate as that of the earth. In a moving vehicle the rotation of the platform would have to be at a greater or lesser rate depending on the direction of motion of the vehicle.

In addition to the stable platform, a complete inertial navigation system makes use of accelerometers. In its simplest form an accelerometer consists of a spring-mass system as shown in Fig. 9.23. By Newton's law a force is needed to accelerate the mass. Whenever the frame of the

device is accelerated therefore, the spring deflects until it generates enough force to accelerate the mass at the same rate as the frame. The deflection of the spring, which may be measured by one of the standard pickup devices discussed in Chap. 5, is a direct measure of acceleration. More sensitive accelerometers may be constructed by mounting a gyroscope such that any acceleration results in a torque on its input axis (see Fig. 9.27 for a possible arrangement). No matter what form the accelerometer takes, it should be clear that it measures acceleration in only one axis. This is really a consequence of the fact that acceleration is a vector

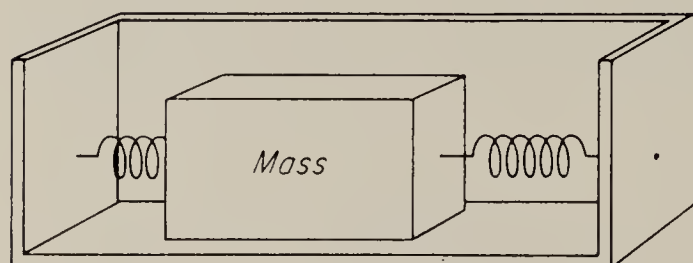


FIG. 9.23. Simplified diagram of an accelerometer.

quantity. It should also be clear that an accelerometer cannot very well distinguish between the acceleration of gravity and other accelerations. Thus, if an accelerometer is to be made insensitive to gravity it must be mounted on an accurately horizontal platform.

Suppose now that we mount three accelerometers on a stable platform on a moving vehicle. The stable platform must be able to maintain accurately the north-south direction, and it should be exactly horizontal. The three accelerometers may then be mounted at right angles to each other such that one of them measures vehicle acceleration along a north-south axis, another measures accelerations along the east-west axis, and the third measures accelerations in the up-and-down direction. If the output of each accelerometer is integrated, the three component velocities are obtained, and if these velocities are again integrated, the position of the vehicle relative to the starting point is obtained. It is clear that if any accuracy is to be realized by this system all parts must be extremely precise. In particular, if the stable platform is not exactly horizontal, the north-south and the east-west accelerometers detect a component of the acceleration due to gravity and the output will therefore be erroneous.

The accelerometer output may be used to maintain the platform horizontal. Qualitatively it is easy to see why this should be so. Assuming that the platform was originally horizontal, and that the coordinates of the vehicle relative to the earth were known, the integrated accelerometer output should indicate exactly where the vehicle is at any later time. Hence it should be possible to compute what the correct platform position should be, and to apply the proper corrections to the torque motors of the gyros.

To illustrate the process of maintaining the platform horizontal somewhat more quantitatively, consider the simplified one-dimensional configuration of Fig. 9.24a. The angle θ_0 represents the angle made by a true vertical with an inertial reference. The indicated vertical, which is a

line drawn perpendicular to the table, makes an angle θ_t with the reference line. Suppose first that there is initially no error, i.e., $\theta_v = \theta_t$, and let the vehicle accelerate with a linear acceleration a . If the accelerometer has a gain constant K_a , then it will deliver an output voltage $K_a a$. As a result of the acceleration the vehicle moves along the surface, and therefore the angle of the true vertical with respect to the reference changes.

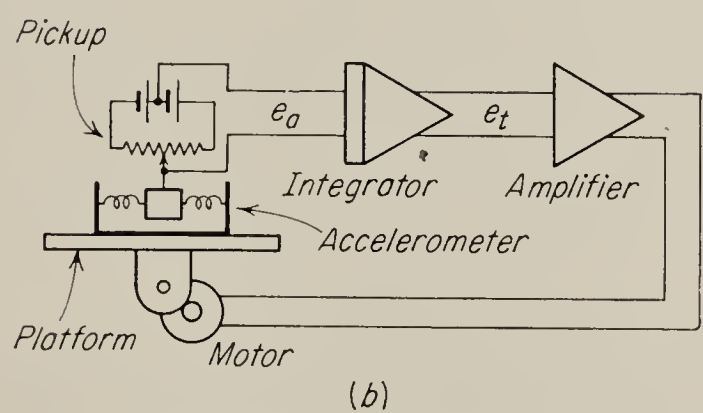
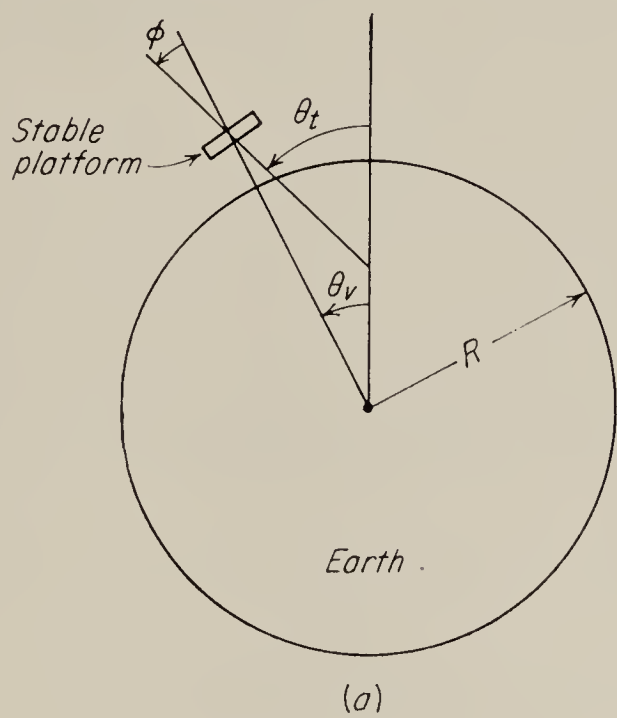


FIG. 9.24. The inertial navigation problem: (a) coordinate system; (b) single-axis block diagram.

If R is the radius of the earth, the acceleration of the true vertical with respect to the inertial reference is

$$\ddot{\theta} = a/R \tag{9.28}$$

A torque must therefore be applied to the gyro mounted on the table such that the table angle θ_t also accelerates this amount. If HIG gyros are used, a fixed voltage applied to the torque motor results in a fixed velocity of rotation of the table. Therefore, we must integrate the accelerometer output once to obtain the angular velocity of the true vertical and apply the output from the integrator to the torque motor. If the gain of the integrator is K_i and if the ratio of table angular velocity to gyro input

voltage is K_t ,

$$\dot{\theta}_t = K_t K_i \int K_a a dt$$

or

$$\ddot{\theta}_t = K_t K_i K_a a \quad (9.29)$$

Using Eq. (9.28) gives

$$\ddot{\theta}_t = K_t K_i K_a R \ddot{\theta}_v \quad (9.30)$$

Since the object of the torquing is to keep the table horizontal we desire that $\theta_t = \theta_v$ and therefore

$$K_t K_i K_a R = 1 \quad (9.31)$$

This is the so-called Schuler tuning condition.¹

It is interesting to observe now what happens as the result of an initial misalignment of the table. Let the angle between the table and the horizontal be $\phi = \theta_t - \theta_v$; ϕ is the error angle. Since the table is not horizontal, the accelerometer reads a component of g ; the acceleration of gravity and its output becomes

$$\begin{aligned} e_a &= K_a(a \cos \phi - g \sin \phi) \\ &\approx K_a(a - g\phi) \end{aligned} \quad (9.32)$$

if ϕ is small. Therefore Eq. (9.30) becomes

$$\ddot{\theta}_t = K_t K_i K_a R (\ddot{\theta}_v - g/R \phi) \quad (9.33)$$

Substituting $\ddot{\theta}_t = \ddot{\phi} + \ddot{\theta}_v$, and applying the Schuler tuning condition [Eq. (9.31)] gives

$$\ddot{\theta}_v + \ddot{\phi} = \ddot{\theta}_v - g/R \phi$$

or

$$\ddot{\phi} + g/R \phi = 0 \quad (9.34)$$

The solution of this differential equation is oscillatory with zero average value and peak amplitude ϕ . Thus the stable platform oscillates about the correct value with a maximum excursion equal to the initial misalignment. Similarly the output from the accelerometer, the velocity integrator, and the distance integrator will oscillate about their correct outputs. We have demonstrated therefore that an initial misalignment does not give rise to an ever-increasing error, but rather to an oscillatory, and therefore bounded, error. Similar results are obtained by considering errors resulting from accelerometer offset, etc.

The angular frequency of oscillation of the table is seen to be $\omega = \sqrt{g/R}$ and the period is $2\pi/\omega = 84.4$ min. This is the Schuler period. It is easy to show² that a compound pendulum having this period continues to point to the center of the earth (i.e., will maintain a true vertical) in spite

¹ M. Schuler, *loc. cit.*

² *Ibid.*

of accelerations along the surface. The stable platform acts very much like such a pendulum.

The discussion of the inertial navigation problem given here has perhaps been quite brief. Nothing has been said about the three-axis problem, and about the necessity of feeding corrections for earth rotation and other factors into the system. For further information on this subject the reader is referred to the literature.¹

PROBLEMS

9.1. A worm-gear differential is shown in Fig. 9.25. One input is the rotation of the shaft X_1 . The second input is the linear displacement of the worm by moving the

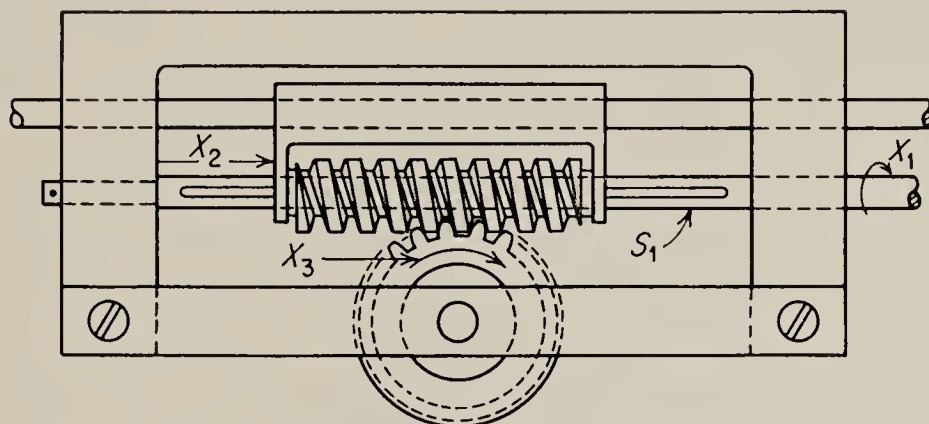


FIG. 9.25. Differential worm gearing.

bracket X_2 . The worm may slide along the spline on X_1 , but it cannot turn with respect to X_1 . The rotation of the gear X_3 is the output. Discuss the possibilities of such a device with respect to limitations on the magnitude of input signals and the possible variations in transfer functions from each input to the output.

9.2. A spiral- and spur-gear differential is shown in Fig. 9.26. Rotation of the shaft X_1 forms one input; the second input is the linear motion of shaft X_2 . The

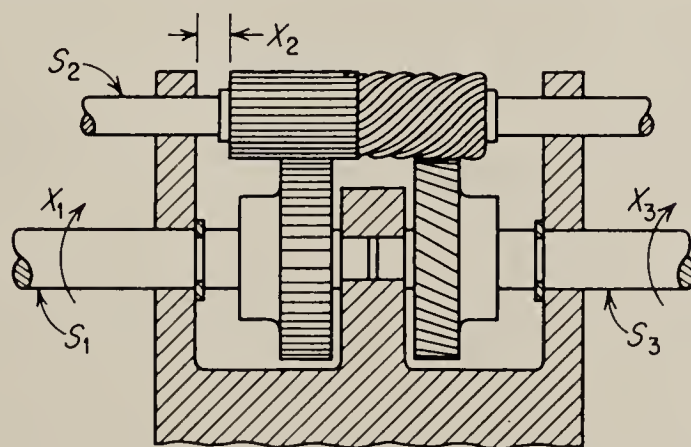


FIG. 9.26. Differential with exactly displaced spiral gear.

output is the angular rotation of shaft X_3 . Discuss the possibilities of such a device with respect to the limitations on input-signal magnitudes and the possible limitations on the transfer functions from each input to output.

¹ N. F. Parker and C. P. Greening, Inertial Navigation, *AGARD, Second Guided Missiles Seminar—Guidance and Control*, pp. 67–86, September, 1956. W. T. Russell, Inertial Guidance for Rocket-propelled Missiles, *Jet Propulsion*, vol. 28, no. 1, p. 17, January, 1958. (This article contains a list of references to other papers.)

9.3. Demonstrate that the nutating frequency of a thin disk spinning in free space is equal to one-half the spin velocity of the disk.

9.4. Write the transfer function of a free gyro from T_z to θ_y .

9.5. Write the transfer function of a rate gyro from T_y to θ_z and construct a block diagram.

9.6. In Fig. 9.27 is shown a simplified diagram of a possible gyroscopic accelerometer. Derive the transfer function.

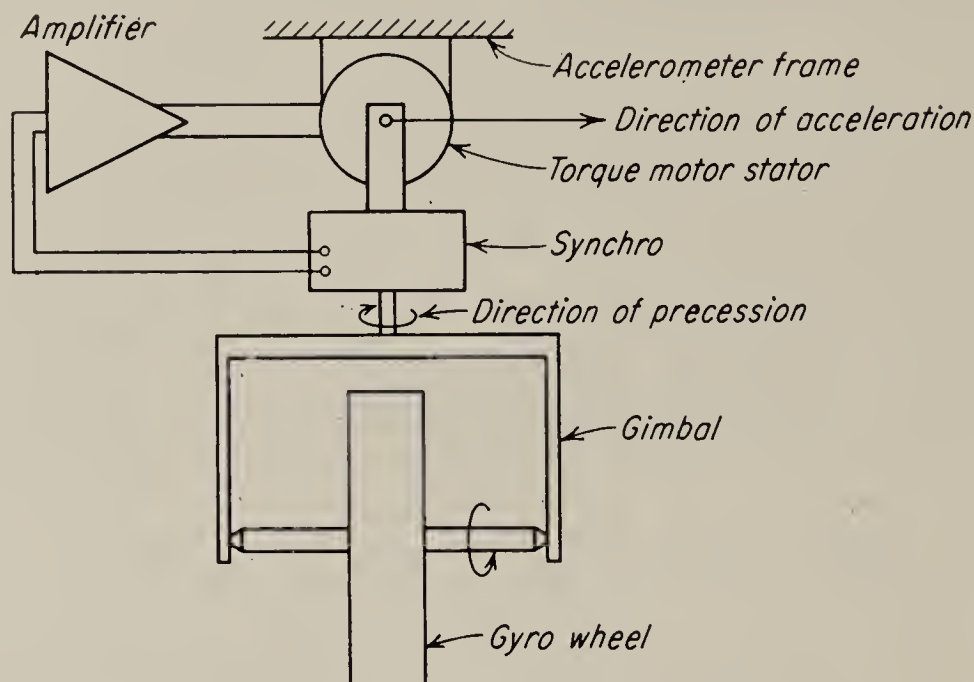


FIG. 9.27. Simplified sketch of a gyroscopic accelerometer.

9.7. In Fig. 9.28 is shown a simplified diagram of a monorail car. Derive the equations of motion and investigate stability.

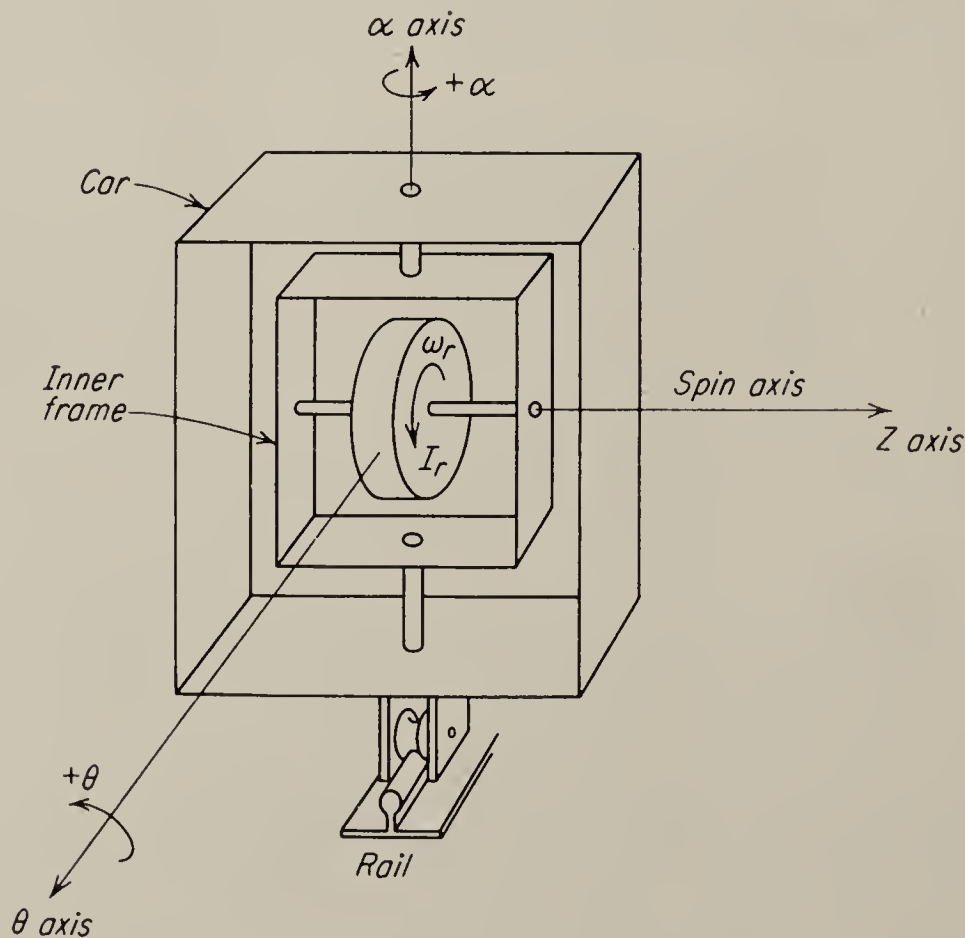


FIG. 9.28. Sketch of monorail car.

CHAPTER 10

PUMP-CONTROLLED HYDRAULIC SYSTEMS

10.1. Introduction. In the general engineering sense, hydraulics is the study of the statics and dynamics of fluids, both liquid and gas. For the control engineer the term has a more restricted meaning. *Hydraulic* is reserved for incompressible (or almost) fluids or liquids, and usually the liquids are restricted to the various types of hydraulic oil, although occasionally it is convenient to use a liquid being processed in the control system itself. The qualification “almost” on the description incompressible is necessary because under certain conditions the compressibility of the liquid must be considered. The intent, however, is to differentiate “hydraulics” from “pneumatics,” in which compressibility is a major consideration.

Hydraulic control systems have several unique advantages. A hydraulic system may be designed to deliver high power to elements separated by some distance, such as throughout an aircraft. Hydraulic systems are thus better for some purposes than mechanical systems, which must be localized; of course, electric systems are superior to all other methods from this aspect. The actual power element or hydraulic motor can be made much smaller physically than an equivalent electric motor. A volume factor of 20:1 for units of about 5 hp may be cited. Owing to this size reduction, the time constant of the hydraulic motor will be much smaller than the time constant of the equivalent electric motor. Hydraulic elements are also more rugged than electric components and are not usually susceptible to noise pickup from shocks and vibration as are vacuum tubes.

Hydraulic systems have their disadvantages as well, of course. Hydraulic lines are not so flexible as electric wire. Hydraulic systems are somewhat more sensitive to environmental temperature variation than electric systems. Hydraulic systems require fairly elaborate power supply and pressure-regulating devices, and finally, in case of damage to one part of the system, there is danger of the entire supply of hydraulic oil being lost.

10.2. Basic Types of Hydraulic Control Systems. There are two basic classes of hydraulic control systems. The first of these classes is the

pump-controlled system, which usually consists of a single variable-stroke pump and a fixed-stroke motor. Control of the motor is exercised by varying the amount of oil delivered by the pump. This is accomplished by mechanically changing the pump volume or stroke, as shown in Fig. 10.1. The pump-controlled system is quite simple and is used when there is only one motor to be controlled. As shown in Fig. 10.1 the motor is the usual rotary form. In certain applications a linear actuator is required.

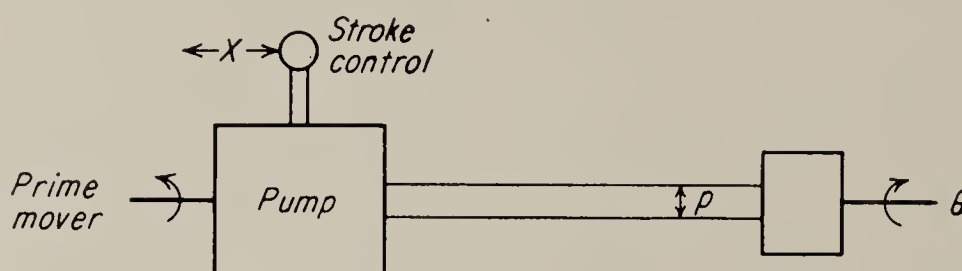


FIG. 10.1. A pump-controlled hydraulic system.

For installations with several motors or actuators that must be independently controlled it is usually more economical and compact to employ one large pump and several motors that are individually controlled by valves. This is the second basic class of hydraulic control systems, the *valve-controlled system*. Although fast-acting, accurate servo valves are expensive, they result in a lighter, more flexible system, and when the relatively inexpensive fixed-stroke pump plus pressure regulator is compared to the several expensive variable-stroke pumps required by independent pump-controlled systems, the total cost of valve-controlled operation is usually less.

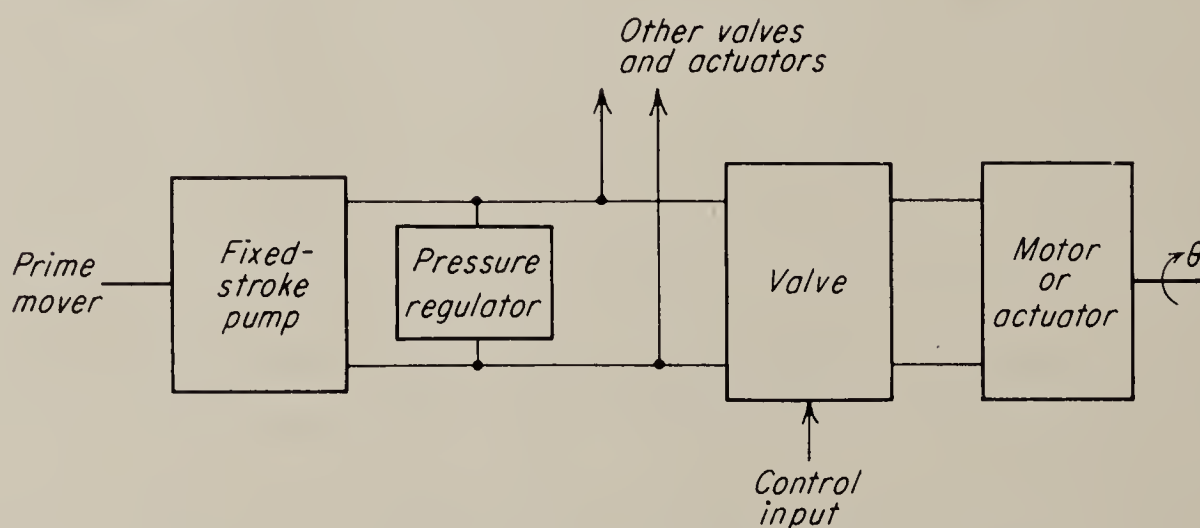


FIG. 10.2. A valve-controlled constant-pressure hydraulic system.

There are two basic types of valve-operated systems. In the first type, the *constant-pressure system* shown in Fig. 10.2, some form of pressure-regulating device is used with the pump to provide a constant pressure supply. When the valve is in the closed position, there is no oil flow from the supply, or, at most, only a small leakage flow. Valves used in this type of operation are referred to as closed-center valves. The second type of system is the *constant-flow type*. In this type, an open-center

valve is used which permits the hydraulic fluid to pass directly to the sump when the valve is in the off position. Operation of the valve diverts part of the fluid away from the sump to the actuator, and in the fully on position all the fluid goes to the actuator. In this type of system the supply pressure is low whenever the valve is in the off position but rises as oil is diverted to the actuator. No pressure regulator is required except possibly as an emergency relief valve, and since fixed-stroke hydraulic pumps are inherently constant-flow devices, neither is any flow regulator required.

If a fixed-stroke pump is used with a pressure regulator in the constant-pressure system, the power delivered by the pump must be essentially constant, since both the flow and pressure are constant. Hence this system is very inefficient, particularly if the duty cycle is such that relatively large peak loads occur only occasionally. Except for the power that is delivered by the hydraulic motor or actuator to its load, all the power delivered by the pump is eventually converted into heat, which raises the oil temperature. Hence fairly elaborate cooling is usually required in this type of system to maintain the oil temperature below its permissible maximum.

Although the constant-flow system does not suffer from this particular disadvantage, it is not used as commonly as the constant-pressure system. This is due to the fact that the control of the actuator is not so positive with this type of system as with the constant-pressure type. It might be noted incidentally that in the constant-flow system the load is not locked rigidly when the valve is in the centered position. Depending on the application, this may or may not be an advantage.

In this chapter we shall consider the pump-controlled hydraulic control system and the various components employed. In the following chapter the valve-controlled system and its components will be discussed.

10.3. The Pump-controlled Hydraulic System. The basic diagram of the pump-controlled system is shown in Fig. 10.1. The most common type of variable-stroke pump used in a system of this type is the piston pump, shown in Fig. 10.3. The prime mover turns the shaft to which the cylinder block is attached. The piston-return plate, or *wobble plate*, is fixed. As the cylinder block rotates, it carries with it the pistons which bear on the wobble plate. The fluid-valving plate is positioned in such a manner that the cylinder volume is increasing as the piston passes the inlet port and decreasing as the outlet port is passed. Thus the fluid that was taken in through the inlet port is expelled through the outlet port.

The flow rate of the piston pump can be changed by changing the angle of the wobble plate with respect to the cylinder block. This adjustment can be made mechanically, hydraulically or electrically. The ease of this adjustment and the fact that maximum flow rate is not affected by pro-

viding for this adjustment make the piston pump the most widely used high-performance variable-stroke pump. In standard piston pumps, considerable reaction force is experienced at the lever that controls the wobble-plate angle. This reaction force is usually in such a direction as to reduce the flow to zero, although it is possible by offsetting the pivot to neutralize this force for a particular load and indeed to reverse the net force. Owing to this reaction force, it is necessary to provide a stage of power amplification to operate the pump stroke control if the stroke is to be controlled electrically. This stroke control could be of the type discussed in Sec. 11.1, for instance.

The fluid is passed from the outlet port of the variable-stroke pump through the high-pressure line to the motor. Any of the various types of hydraulic pumps theoretically work equally as well when operated as

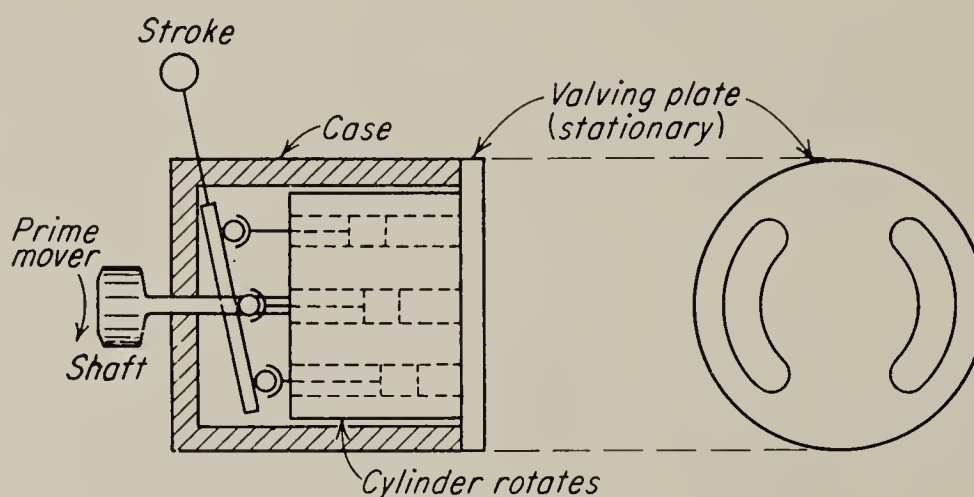


FIG. 10.3. Basic variable-stroke piston pump.

motors; however, motors are almost always built with a fixed stroke. Thus, a piston type of motor has a wobble plate fixed at a particular angle (usually 30 to 40°). The high-pressure fluid is admitted by the valving plate to a cylinder and forces the piston back. This can be accomplished only by the cylinder's turning with respect to the stationary wobble plate, thus turning the output shaft. The cylinder which has been filled by the fluid is then carried around to the exhaust port and the fluid expelled to the low-pressure line by the returning piston. Other types of pumps and motors that are discussed in a later section can also be employed in such a system with varied degrees of success.

All hydraulic transmissions must have some sort of make-up system, to replenish the fluid lost by leakage, and must also have an overload relief valve, for protection in case of load seizures, etc. A typical replenishing system and relief protection are shown in Fig. 10.4. In the usual case of bidirectional operation, either line may be the high-pressure line; thus a relief valve should be placed in both lines. Similarly, replenishing is usually provided in both lines to provide for the possibility of operation in either direction for long periods of time. Care must be taken that the

fluid picked up from the sump is not aerated. Defoaming agents are sometimes added to the hydraulic fluid in order to reduce its compressibility and to improve hydraulic efficiency. The replenishing feature adds a nonlinearity to the operation of the system since it places a lower limit below which the return-line pressure may not drop.

The relief valve also contributes a nonlinearity to system operation. The pressure in the high-pressure line is proportional to the developed torque on the motor. Thus a relief valve sets an upper limit on the torque that can be produced by the motor. In the case of a pure-inertia load this sets an upper limit on the motor acceleration and, in the case of a viscous-damping load, an upper limit on the motor velocity.

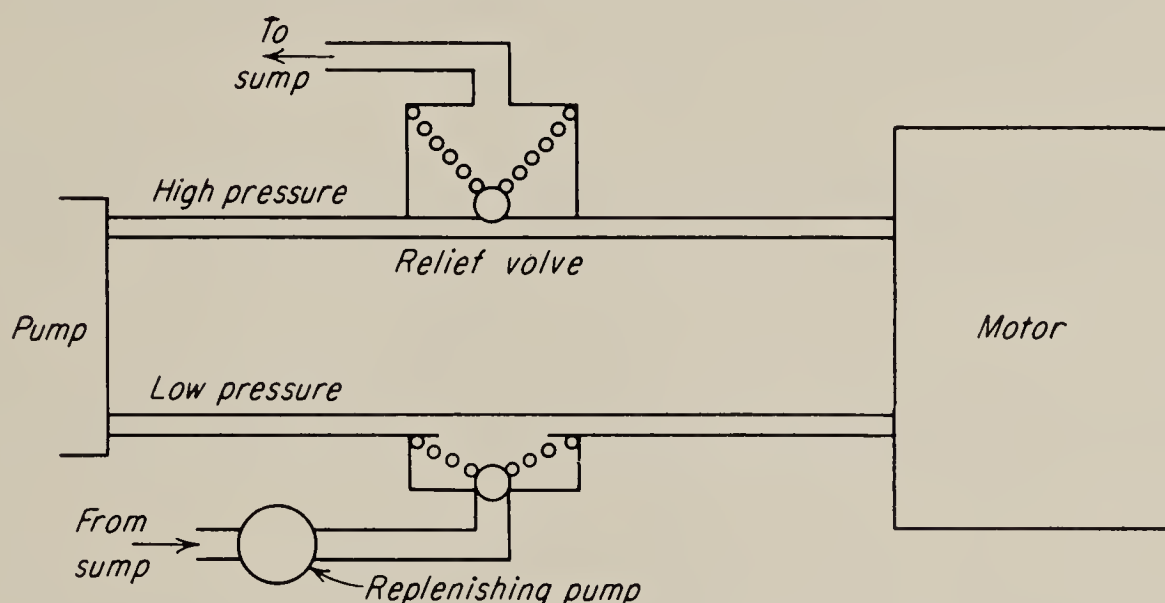


FIG. 10.4. Hydraulic transmission with replenishing valve and overload relief valve.

10.4. Analysis of the Pump-controlled System. The delivery rate of an ideal piston pump can be calculated as the product of the volume swept out by a piston per revolution times the number of pistons times the speed of revolution.

$$F_p = xAcn = k_px \quad (10.1)$$

where $F_p \triangleq$ pump-flow rate, in.³/sec

$x \triangleq$ length of stroke, in.

$A \triangleq$ cross-sectional area of piston, in.²

$c \triangleq$ number of cylinders

$n \triangleq$ speed, rps

Ideally all the pump flow passes through the motor and is converted to output shaft speed. Hence the speed of the motor shaft is theoretically directly proportional to the pump-stroke position x .

When there is no load on the motor shaft, there is no back pressure in the motor, and all the fluid flows through the motor in the proper channels. When there is a load on the motor shaft, however, there must be a pressure built up by the hydraulic flow to overcome this resistance to free

movement. The free flow of fluid in the normal motor channels is impeded, and fluid leaks through gaskets and around end plates. It is usually assumed that the leakage flow is proportional to motor pressure or, since it will be shown that pressure is proportional to developed torque, to developed torque in the motor.

$$F_L = K_L \hat{T}_d \quad (10.2)$$

Thus the pump flow does not all go into useful work in the motor, and as we shall see below, an additional portion of flow must be accounted for.

To show that the pressure across the motor is proportional to developed torque, we recognize that the angular displacement of the motor shaft is directly related to the amount of oil (q_m) forced through the proper channels in the motor. Thus

$$\hat{q}_m = K_m \hat{\theta} \quad (10.3)$$

where K_m is the motor displacement constant, and θ is the angular position of the motor shaft. Differentiating Eq. (10.3) gives

$$s\hat{q}_m = \hat{F}_m = K_m s\hat{\theta} \quad (10.4)$$

By the law of conservation of energy the hydraulic power must be equal to the mechanical power developed by the motor. Thus,

$$p \frac{\text{lb}}{\text{in.}^2} \times F \frac{\text{in.}^3}{\text{sec}} = T_d \dot{\theta} \quad \frac{\text{in.-lb}}{\text{sec}} \quad (10.5)$$

Since from Eq. (10.4) we have

$$F_m = K_m \dot{\theta} \quad (10.6)$$

therefore

$$T_d = K_m p \quad (10.7)$$

The back pressure built up by the load not only causes leakage flow in the motor and in the pump as well, which is accounted for by the term F_L , but also causes the fluid in the lines to compress. The compressibility of the hydraulic fluid is partially due to the minute quantities of air entrained in it and to the expansion of the tubing with pressure, but the major factor is the oil itself, which is compressible, as is any solid or fluid. Usually the compressibility is a second-order effect except when the system operates at very high pressures and when there is a fairly large amount of oil in the system. Volume compressibility is essentially proportional to pressure. Thus the compressibility flow is proportional to the rate of change of pressure

$$\hat{F}_c = K_c s\hat{p} \quad (10.8)$$

We can calculate the coefficient of compressibility, K_c , by assuming that the motor is stalled, so that F_m is zero, and that the leakage flow F_L is also zero. The pump pulls a volume of oil out of the low-pressure line

and forces it into the high-pressure line. The compression volume in the high-pressure line is

$$\Delta q_1 = \frac{V_1 \Delta p_1}{\gamma} \quad (10.9)$$

where q = oil volume, in.³

p = pressure, psi

V = volume of pipe, in.³

γ = bulk modulus of fluid, psi

Δp = change in pressure in line, psi

and where the subscript 1 refers to the high-pressure line.

Similarly the compression volume in the low-pressure line is

$$\Delta q_2 = \frac{V_2 \Delta p_2}{\gamma} \quad (10.10)$$

The two compression volumes will be equal in magnitude and opposite in sign, because the fluid pumped from one line is identical to the fluid pumped into the other. Hence

$$\Delta q_1 = -\Delta q_2 = \Delta q_c \quad (10.11)$$

Then if

$$V_1 = V_2 = \frac{V}{2} \quad (10.12)$$

where V is the total oil volume under consideration, we may rewrite Eq. (10.9) as

$$\Delta q_c = \frac{V}{2\gamma} \Delta p_1 \quad (10.13)$$

Since the two volumes are the same, and if the replenishing valve does not open, we have also that

$$\Delta p_1 = -\Delta p_2 = \frac{\Delta p}{2} \quad (10.14)$$

where Δp is the pressure difference between the two lines. Then

$$\frac{\Delta q_c}{\Delta p} = \frac{V}{4\gamma} = K_c \quad \text{in.}^5/\text{lb} \quad (10.15)$$

where K_c is the compressibility coefficient. The bulk modulus of standard hydraulic fluids is usually about 2.5×10^5 psi; hence the compressibility coefficient is about 10^{-6} in.⁵/lb per cubic inch of fluid. Newton¹ points out that, at atmospheric pressure, air makes up about 0.2 per cent by volume of the fluid bulk, causing a considerable increase in K_c .

¹ G. C. Newton, Speed Transmissions as Servo Motors, *J. Franklin Inst.*, vol. 243, p. 458, 1947.

Zweig¹ has shown that the compressibility coefficient is increased about 25 per cent in a standard 1/2-in. copper hydraulic line by expansion of the tube walls under pressure.

The theoretical pump flow F_p is absorbed in these three component flows and may thus be expressed by

$$\hat{F}_p = \hat{F}_m + \hat{F}_c + \hat{F}_L \quad (10.16)$$

Using the flow relations above and given a knowledge of the load, it is possible to derive a transfer function from pump stroke x to output shaft position θ .

10.5. Transfer Function of the Pump-controlled System. In order to derive the transfer function of the pump-controlled system, it is necessary to determine the relation for the shaft load. Let us assume for generality that the load consists of inertia, viscous damping, and an arbitrary load torque. The relation in terms of developed torque and shaft position will then be

$$\hat{T}_d = (Js^2 + Bs)\hat{\theta} + \hat{T}_L \quad (10.17)$$

where J is the moment of inertia, B is the coefficient of viscous damping, and \hat{T}_L is the arbitrary load torque.

Substituting the developed relations for the various flows into Eq. (10.16) and then eliminating the extraneous variables by the use of Eqs. (10.17) and (10.8) yields

$$\theta = \frac{K_p \hat{x} - [K_L/K_m + (K_c/K_m)s]\hat{T}_L}{s[(JK_c/K_m)s^2 + (JK_L/K_m + BK_c/K_m)s + K_m + BK_L/K_m]} \quad (10.18)$$

As we would expect, the gain of the system is determined primarily by the pump constant K_p . The free s in the denominator indicates an integration. The presence of compressibility combined with other parameters provides an additional degree of freedom to the system and thus allows the possibility of underdamped roots or resonance at some critical frequency. In small, localized systems, K_c approaches zero, and the K_c terms may be neglected. In systems with long lines between pump and motor, it may not be possible to neglect the pressure drop in the line due to flow, as was done in this analysis. The problem of the hydraulic transmission line is discussed in some detail in a later section (Sec. 10.11).

10.6. Other Pump and Motor Types. Hydraulic pumps may be divided into two classes, positive-displacement pumps and turbopumps. Generally speaking, only positive-displacement pumps are used in control-system applications. Pumps are also classified as radial flow or axial

¹ F. Zweig, "Compressibility Effects in Hydraulic Transmission Lines," Yale Univ. Dept. Elec. Eng. Tech. Rept., June, 1950.

flow for purposes of description. Actually there seems to be no limit to the configurations proposed for pumps and motor designs, and those discussed here are merely representative. The various problems of leakage, wear, etc., appear in all designs, and in general a device may be used interchangeably as a motor or a pump.

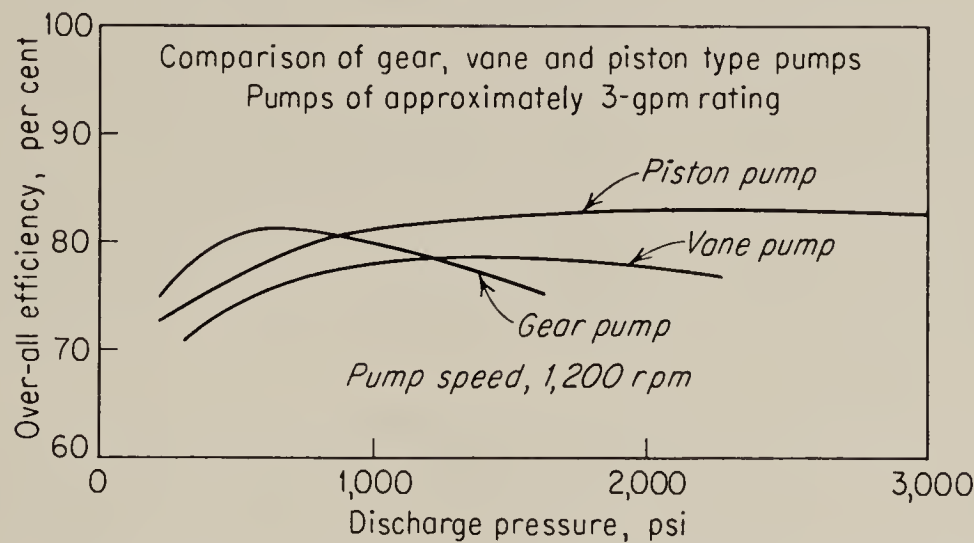


FIG. 10.5. Over-all efficiency versus discharge pressure for the three most common types of pumps. (From Kaay et al.)

The three most common types of positive-displacement pumps are the piston pump, which was used as an example in the pump-controlled system, the vane pump, and the gear pump. Piston pumps may be adapted for variable delivery with no loss in efficiency and delivery, while gear pumps are essentially fixed-displacement devices. Vane pumps lie between these two extremes. Figures 10.5 and 10.6 show sample test data for the

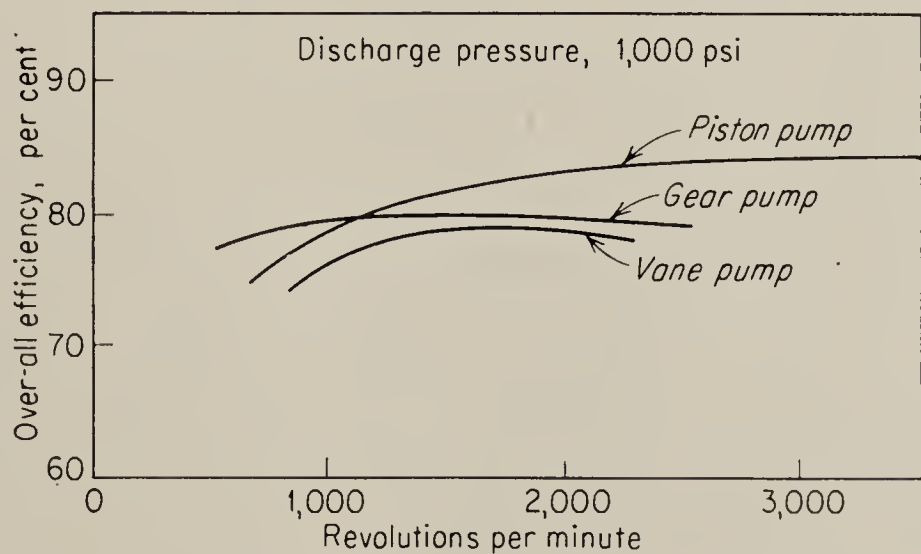


FIG. 10.6. Efficiency versus operating speed for the three most common types of pumps. (From Kaay et al.)

three types of pumps at various pressures and operating speeds.¹ The pumps used to obtain the test data were rated at 3 gpm and designed for the same type of service. The three types of pumps will be discussed in detail in the following sections.

¹ H. A. Vander Kaay, R. J. Murphy, and W. Traut, Hydraulic Pumps, *Product Eng.*, vol. 26, no. 7, p. 132, 1955.

Basically, all hydraulic pumps may be used as hydraulic motors, and usually a hydraulic system consists of a matched set, since the factors such as speed and pressure that dictated the choice of the pump type are equally operative in the choice of the motor. Usually the stroke of a hydraulic motor is fixed, and the motor is controlled by controlling the fluid flow. Surprisingly, the simplest of the pumps, the gear pump, has in the past been unsuccessful as a motor. The hydraulic unbalance wedges the gears at standstill and reduces the available starting torque of the standard gear motor to an unacceptable level.

10.7. Piston Pumps. Piston pumps are the most complex and therefore the most expensive of the three main types of positive-displacement pumps. The basic piston pump was shown in Fig. 10.3, and the operation has been described.

The pistons and cylinder block, while they must be held to very close tolerances, are not the source of most of the leakage flow or of the manufacturing difficulty. The clearance between cylinder block and valving plate is the major source of design problems in most models.

Early piston pumps had only a few pistons, and this resulted in pulsating fluid delivery. By increasing the number of pistons and decreasing their diameter, the same rate of flow can be maintained while the pulsations are minimized.

The power delivered by a hydraulic pump is the product of pressure and flow. The flow is equal to the product of the displacement of the pump per revolution times the speed of the pump. The weight and size of the pump are primarily functions of the pump displacement. Thus, if the operating pressure and the pump speed are increased and the displacement decreased, the output power can be maintained while the size and weight of the pump are reduced. These factors also reduce the size and weight of other components in the system. Pressures and pump speeds have thus shown continuous increases in applications such as use in aircraft, where weight is of primary importance. Piston pumps have been used at speeds as high as 6,000 rpm and pressures up to 8,000 psi without exceeding critical fluid velocities, etc.¹ It will be noted from Figs. 10.5 and 10.6 that the piston pump is superior to the other pump types at the higher operating speeds and pressures.

10.8. Vane Pumps. The vane pump is a considerably simpler mechanical device than the piston pump. In the fixed-displacement vane pump with two inlet and two outlet ports, the hydraulic forces are balanced, and the bearings experience essentially no hydraulic reaction force. A fixed-displacement vane pump is shown in Fig. 10.7. Low-pressure fluid is picked up by a vane as it passes an inlet port and is forced out the outlet port at high pressure. By proper design the displacement of the vane

¹ *Ibid.*

slots can be made to contribute to the output. This may represent an increase of 10 to 15 per cent in the displacement of the pump.¹ It is only in the larger sizes that the additional porting required for the inlet and the outlet flow is worth while, however. It should be noted that it is not leakage flow that is displaced but fluid introduced into the vane slot specifically for this purpose.

Centrifugal force is usually sufficient to maintain contact between the vanes and the stator, although the vane slots are sometimes pressurized to eliminate an adverse pressure differential, or the vane may be spring-loaded. Ideally the vane is in line contact with the stator, which is not so desirable as the surface contact of the pistons in the piston pump. The

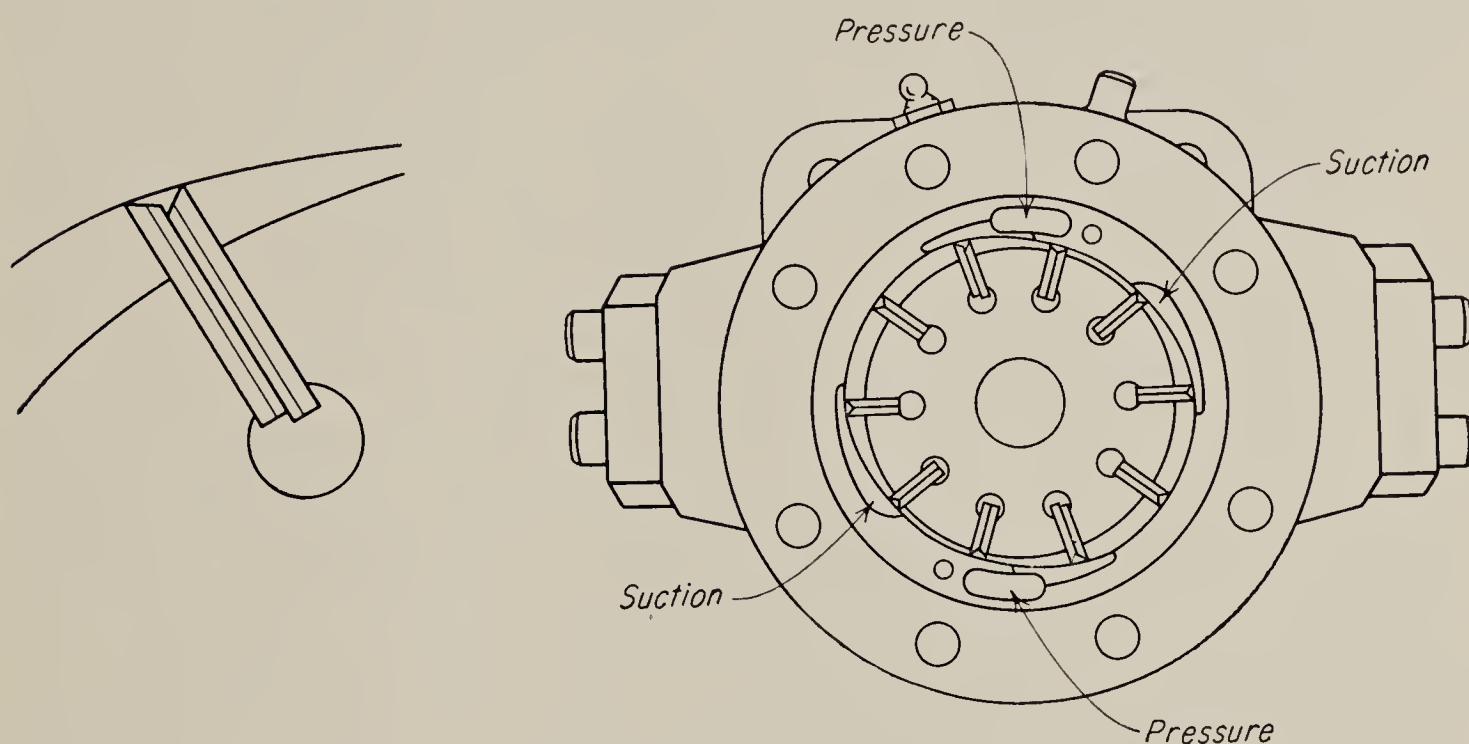


FIG. 10.7. Fixed-stroke vane pump with two inlet ports and two outlet ports.

vanes are the weak point in vane pumps, and the problem of chipping and breaking has not been completely solved. However, it will be noted here that the gradual wear of the vane edges during operation does not increase the leakage flow as wear on the pistons would.

The reciprocating mass in the vane pump is quite small as compared with the piston pump, and the flow of fluid is smoother. Along with these factors the simplicity of construction contributes to the quiet, smooth operation and the long life of the vane pump.

Vane pumps can be constructed with a variable stroke. A simplified sketch of a variable-stroke vane pump is given in Fig. 10.8. When a variable stroke is provided, only one inlet and outlet port may be used. Because of the cam ring and end clearances required for the variable-stroke pump, its volumetric efficiency will be lower than a fixed-displacement pump of the same size. In addition, the delivery volume of a variable-displacement pump is one-half that of the fixed-stroke type.

¹ *Ibid.*

The variation in displacement may be made by changing the eccentricity, by sliding the cam ring, or by rotating the cam ring with respect to the ports. The variable-displacement pump is unbalanced hydraulically, and the bearings and cam ring must be designed to support the hydraulic

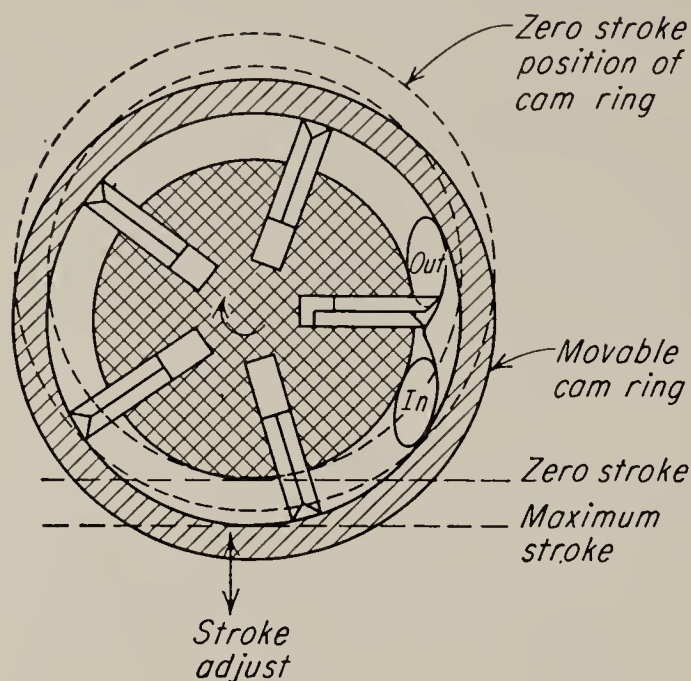


FIG. 10.8. Variable-stroke vane pump. Note that only one inlet and one outlet port may be used.

force produced when the pump operates against pressure. For these reasons the variable-displacement vane pump is not usually designed for a flow of more than 70 gpm or a pressure of over 1,000 psi.¹

10.9. Gear Pumps. Gear pumps have fewer moving parts than the other two common types of hydraulic pumps and are the easiest to manufacture. They are simple, rugged, efficient, and low in cost. Figure 10.9 shows a typical gear pump and the direction of fluid flow. One of the gears is driven from the prime mover, and the other gear idles. The fluid is caught

up in the space between the rotating teeth and the housing and is carried to the outlet under pressure. It will be seen that the gear pump is intrinsically a fixed-displacement device.

Losses in the gear pump may occur in several ways. Leakage can occur between the gears and the housing. This can be held to a minimum by proper bearing design which permits the clearance between housing and gears to be held within close limits. Another source of leakage is the space between the gears and the end plates. Usually these end plates are fixed, with a clearance between them and the gears of several ten-thousandths of an inch. Occasionally the end plates are made free to slide axially and are held tightly to the gears with pressure springs. This results in higher hydraulic efficiency but also higher mechanical losses. A somewhat more elaborate method of holding the end plates against the gears is by hydraulic pressure. In one design the pressure varies with load flow, thus maintaining minimum leakage with minimum loss in efficiency.

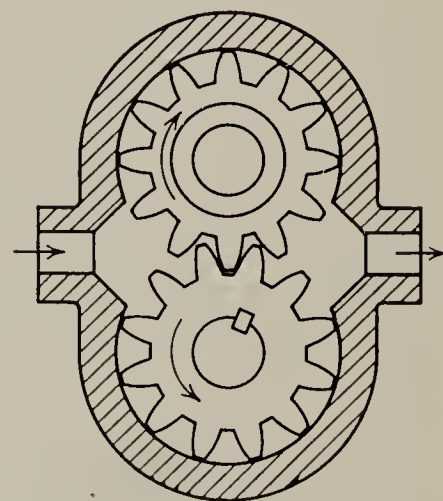


FIG. 10.9. A gear pump.

Simple gear pumps of the type shown in Fig. 10.9 are hydraulically

¹ *Ibid.*

unbalanced, and the gear bearings must be designed to withstand the forces involved. It is possible, however, to design a gear pump with two inlets and two outlets, thus balancing the hydraulic forces. A disadvantage of the gear pump is its operational noise. Gear pumps are made in wide ranges of displacements and speeds. They may be operated at speeds up to 4,000 rpm and occasionally as high as 9,000 rpm; however they are somewhat more efficient at low speeds, as may be seen from Fig. 10.6.

Several problems are encountered in operating gear pumps at high speeds. First, it is difficult to fill the displacement volume between the teeth and housing completely. Some gear-pump models include a nozzle in the inlet port to increase the fluid velocity as it enters, thus forcing the fluid into the displacement volume. An additional problem in gear-pump operation is the reduction in volumetric efficiency that occurs when fluid is trapped between the gear teeth at the outlet port and carried back to the inlet. One ingenious solution to this problem is shown in Fig. 10.10. The driven gear rides on a stationary shaft. Small ports in the driven-gear teeth allow the trapped fluid to flow into the passage in the stationary shaft. The fluid is then added to the output flow. In addition to the

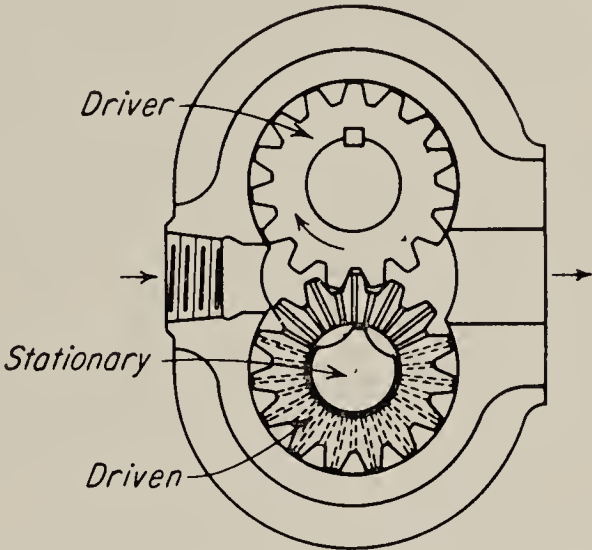


FIG. 10.10. A gear-pump design that recovers the oil trapped between the two gears.

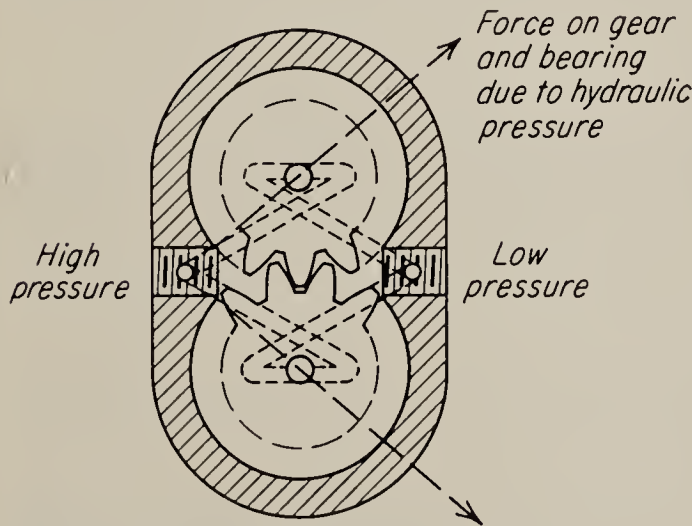


FIG. 10.11. Hydraulic-force compensation of a gear motor. (Courtesy Eastern Industries, Inc.)

increased flow, this method also reduces the fluid pressure between the teeth that tends to separate the teeth and cause bearing wear. While dirt should not be allowed to enter any hydraulic system, the gear pump will experience less damage from contamination than the other forms of pumps, since the particles will be thrown to the outer rim of the gears and be expelled to the system. The standard gear pump is unsatisfactory as a hydraulic motor, because wedging is caused by the unbalanced force from the high-pressure inlet. However, an interesting modification of the gear pump can balance this hydraulic force. In Fig. 10.11 is shown a gear motor with hydraulic balancing. Passages in the case bring high-pressure fluid to bear on the gear shafts at a point designed

to cancel the hydraulic force on the gears. The method is successful in reducing the breakaway pressure required by the motor at starting to 10 to 20 psi, a negligible value.

10.10. The Ball Pump. The ball pump, shown in Fig. 10.12, is basically a modification of the piston design. The balls are forced in and out radially by the cam plate and hydraulic pressure, thus displacing fluid, in the case of a pump, or forcing rotation, in the case of a motor. The inlet and outlet ports are inside the ball race. The stroke may be made vari-

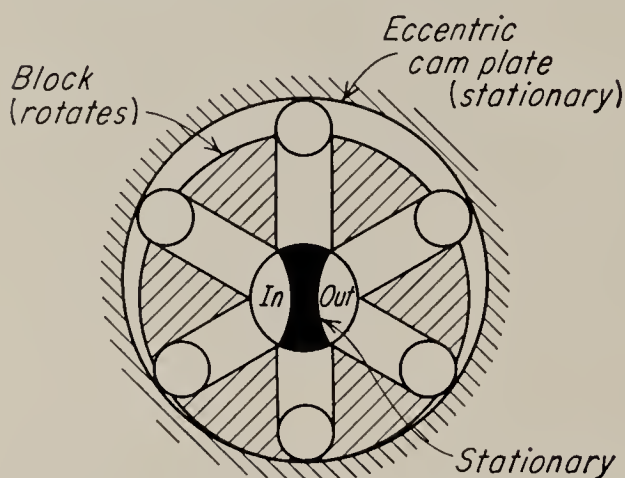


FIG. 10.12. Ball pump.

able by allowing the cam-plate eccentricity to be adjusted. In motor operation, high-pressure fluid is admitted at the port marked in and forces the balls radially outward. This may be accomplished only by rotary motion of the ball race. The ball pump provides a line contact between ball and race, while the conventional piston pump provides a surface contact between piston and cylinder. Thus the problem of leakage

due to ball wear or eccentricity is severe in the ball pump. The radial design has the advantage over the conventional axial piston pump of simplicity and ease of manufacture.

10.11. Hydraulic Transmission Lines. While the effects of very short transmission lines, if they provide adequate cross-sectional area, may be neglected, it is often necessary to consider the pressure drop in a line of moderate length; and in long lines the effect of the dynamic flow charac-

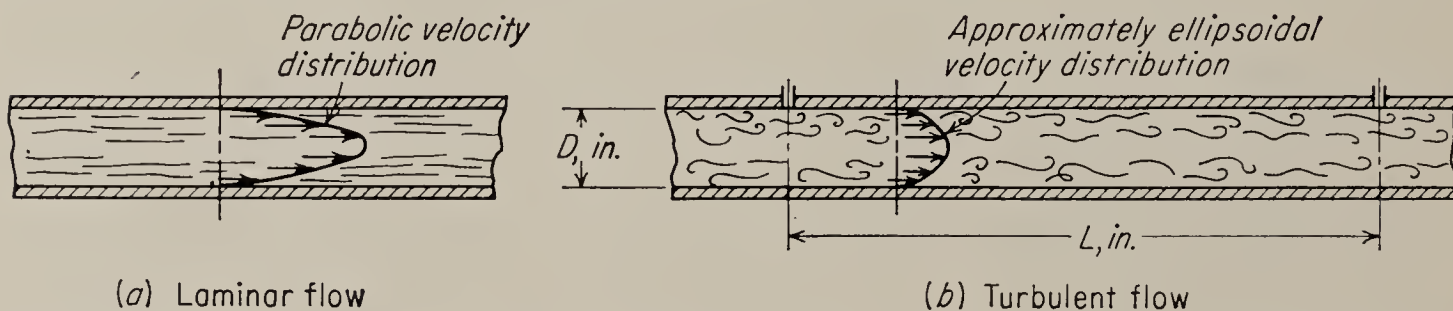


FIG. 10.13. (a) Laminar flow with a parabolic velocity distribution; (b) turbulent flow with approximately ellipsoidal velocity distribution.

teristics upon the transfer function of the control system must also be considered.

Hydraulic flow may be divided into two types: first, viscous flow, or *laminar flow*, and, second, *turbulent flow*. In laminar flow the particles of fluid flow smoothly without jumbling. The flow may be thought of as concentric tubes of fluid sliding on one another, after the manner of an expanding telescope. The innermost tube has the highest velocity, while the tube nearest the boundary is stationary. The tubes of flow do not

exist in turbulent flow. The individual oil particles are constantly inter-mixing. The two types of flow are shown diagrammatically in Fig. 10.13. Under low pressures and for fluids of moderate to high viscosity the flow may be laminar, but as the pressure is increased, the flow becomes turbulent. Flow may be characterized by a dimensionless constant N_R , the so-called Reynolds number, after Osborne Reynolds, whose classic experiments in 1883 defined the types of flow. The Reynolds number is defined as

$$N_R = \frac{DV}{\nu} \quad (10.19)$$

where D = diameter of the fluid section¹, in.

V = mean velocity of flow, in./sec

ν = kinematic viscosity of fluid, in.²/sec

This dimensionless ratio occurs throughout fluid mechanics and is essentially independent of the particular fluid and configuration. Thus it is possible to say that, for a given Reynolds number, flow will be laminar for oil in a pipe or for air around an aircraft wing.

When the flow is increased slowly in a smooth straight pipe through a carefully rounded entrance, laminar flow has been maintained for values of the Reynolds number as high as 50,000. Under normal conditions, however, turbulence commences at Reynolds numbers of 2,400 to 3,000, and if the installation has many bends, fittings, and valves, the flow may be turbulent at N_R 's as low as 1,000.* The pressure drop and resultant energy loss are significantly higher in turbulent flow than in laminar flow. The relationship for flow through a pipe may be given by the Chézy flow formula

$$p_1 - p_2 = \frac{\epsilon \rho F^2}{2gA^2} \quad (10.20)$$

where $p \triangleq$ pressure, psi

$\epsilon \triangleq$ coefficient of energy loss

$\rho \triangleq$ unit density of fluid, lb/in.³

$F \triangleq$ volumetric flow, in.³/sec

$g \triangleq$ acceleration of gravity, in./sec²

$A \triangleq$ cross-sectional area of pipe, in.²

The Chézy relation was developed empirically in 1775, but it may be derived from the law of conservation of energy.² The coefficient of

¹ The concept of the Reynolds number has been found useful throughout the field of hydrodynamics, and the parameter D is in general some characteristic dimension of the system.

* N. M. Sverdrup, *Theory of Hydraulic Flow Control*, *Product Eng.*, vol. 26, no. 4, p. 165, 1955.

² See any text on fluid mechanics, e.g., Dodge and Thompson, "Fluid Mechanics," McGraw-Hill Book Company, Inc., New York, 1937.

energy loss, ϵ , depends on the coefficient of friction and the normalized length of the pipe:

$$\epsilon = f \frac{L}{D}$$

(10.21)

Empirical values of f for smooth pipe as a function of the Reynolds number are given in Fig. 10.14. It will be noted that for laminar flow the

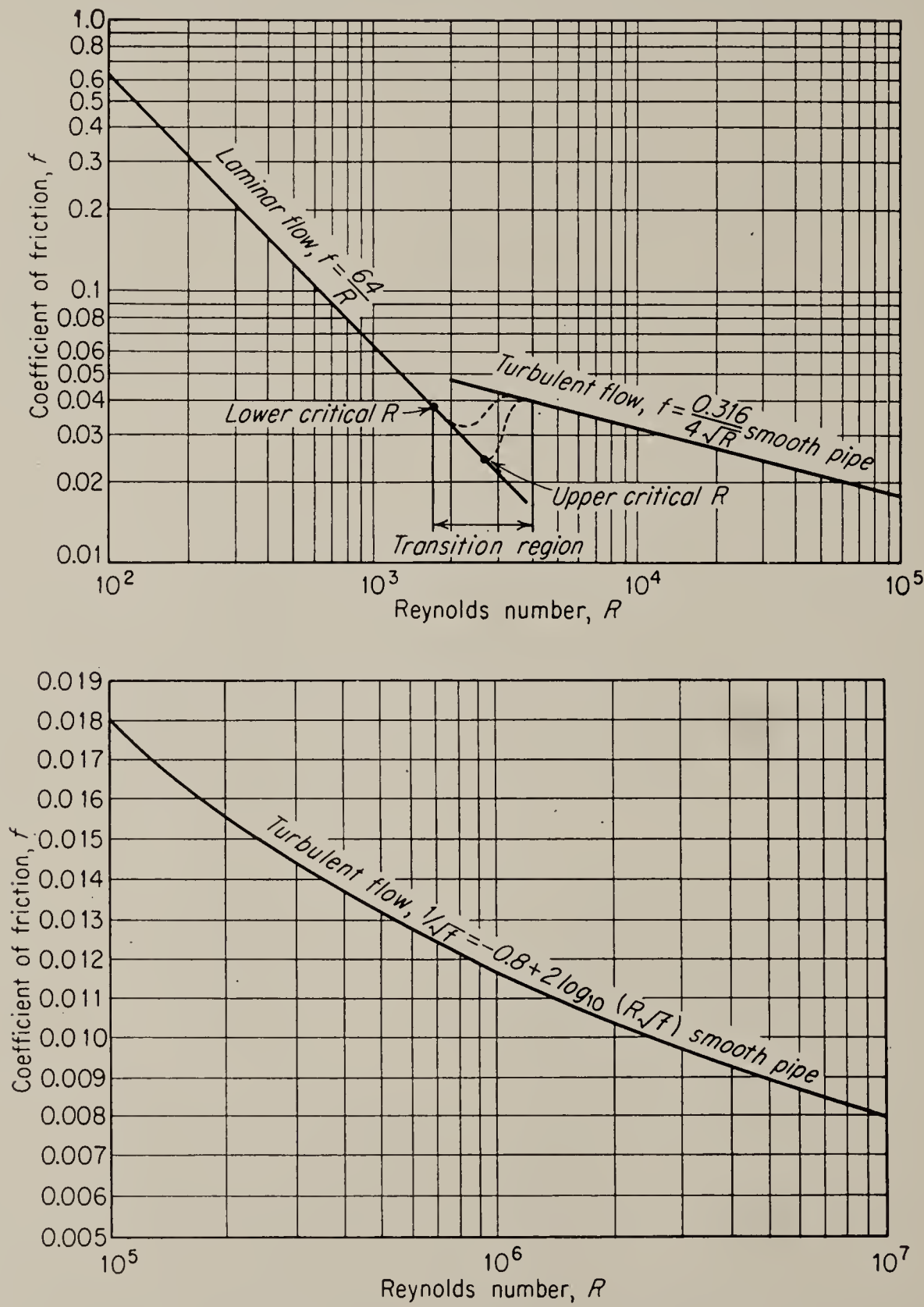


FIG. 10.14. Friction coefficient versus N_R for pipes.

variation of f is linear, and as might be assumed, a relation may be developed for this region that eliminates the need of the diagram and the Chézy formula. This relation is called the Hagen-Poiseuille law for laminar flow and is derived in texts on fluid mechanics.

$$F = \frac{\pi(p_1 - p_2)D^4}{128\mu L}$$

(10.22)

where μ is the absolute viscosity of the fluid. The relationship for flow shapes other than circular pipes may also be derived. The absolute viscosity μ in Eq. (10.22) is related to the kinematic viscosity ν in the definition of the Reynolds number by

$$\nu = \frac{\mu}{\rho g}$$

(10.23)

where μ = absolute viscosity, lb-sec/in.²
 ρ = density of fluid, lb/in.³

Both the absolute viscosity and the density of hydraulic fluids vary with temperature; thus the kinematic viscosity is also a function of temperature. Figure 10.15 shows the density and the kinematic viscosity of

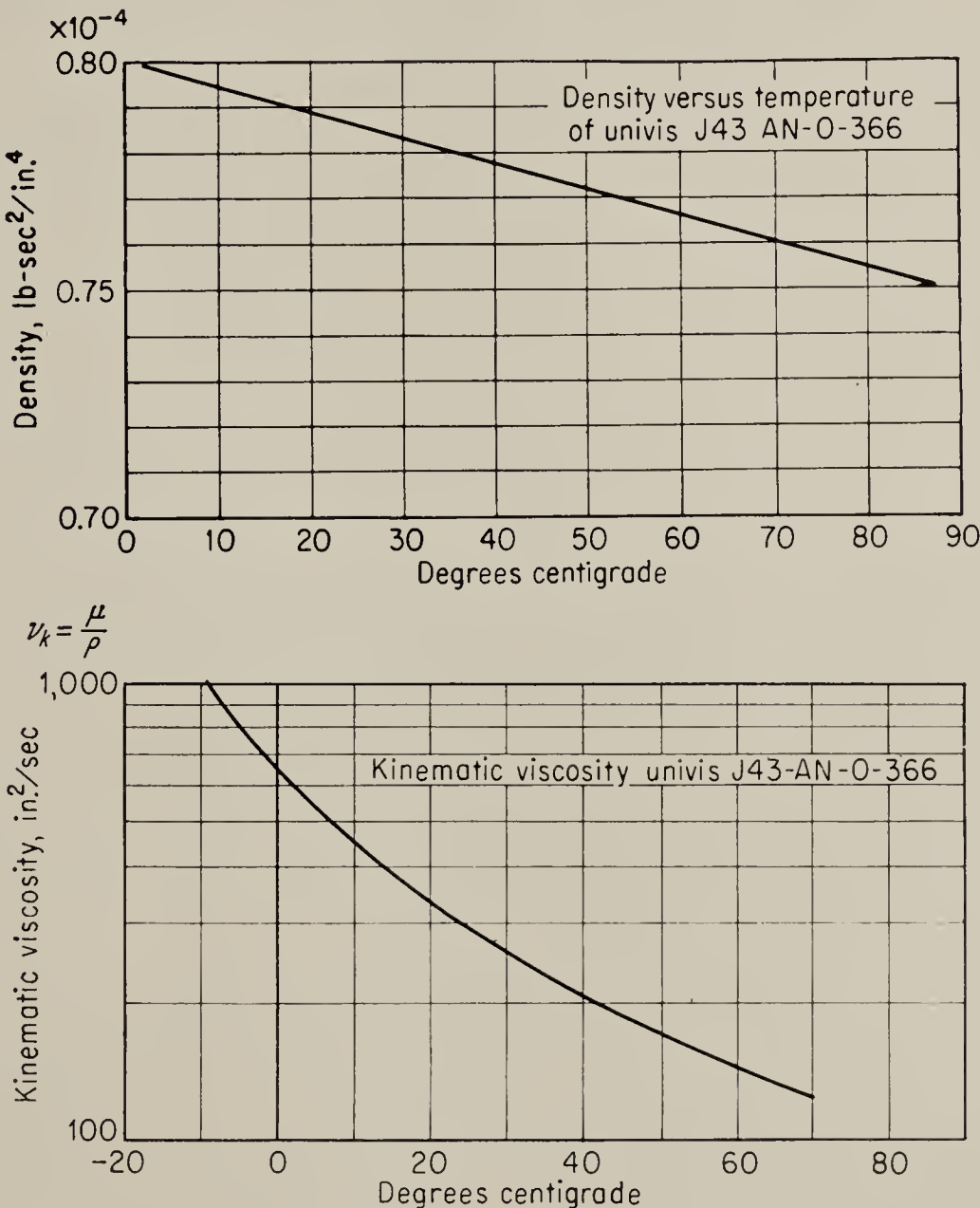


FIG. 10.15. Density (ρ) and kinematic viscosity (ν) of Univis J43 versus temperature. Univis J43 AN-VVO-366, a standard hydraulic fluid, plotted against temperature.

When the hydraulic transmission lines become very long, wave effects and related phenomena must be considered, and the line will then have a

very pronounced effect on the transfer function relating pump stroke to motor speed. Zweig¹ has shown that a hydraulic transmission line behaves in the same general way as an electric transmission line, and that the techniques used for many years in analyzing electric lines may be used also for hydraulic lines. In particular, he has shown that the well-known Smith chart may be used to simplify computations connected with such a line. The criterion for the length of a transmission line is the velocity of propagation of the line. If the time required for a signal to pass down the line is short with respect to the period of the highest-frequency wave that is to be transmitted, the transmission effects may be ignored. As will be shown below, the velocity of propagation in a dissipationless line is given by

$$v = \sqrt{\frac{\gamma g}{\rho}}$$

where γ is the bulk modulus of the oil, psi; ρ is the density in lbs/in.³, and g is the acceleration of gravity. For most hydraulic fluids the velocity is in the neighborhood of 4,200 ft/sec. Hence a 100-ft line is equivalent to one wavelength for a signal frequency of 42 cps and would be considered long at this frequency. It could be considered short at frequencies of less than 4 cps.

The parameters determining the behavior of a transmission line are the surge, or *characteristic impedance* Z_o , and the *propagation constant* $\alpha + j\beta$. The characteristic impedance is the complex ratio of the sinusoidal voltage to current in a wave passing down the line without reflection from the end. The propagation constant is a complex number $\alpha + j\beta$, where α is the attenuation constant expressed in nepers per unit length, and β is the phase shift constant expressed in radians per unit length. It is shown in standard texts on transmission line theory that the velocity of propagation is ω/β , where ω is the signal frequency in radians per second.²

For electrical lines it has been shown that the characteristic impedance is given by

$$Z_o = \left[\frac{R^2 + (\omega L)^2}{(\omega C)^2} \right]^{1/4} / \underbrace{-\frac{1}{2} \tan^{-1} \frac{R}{\omega L}}$$

The attenuation constant α is given by

$$\alpha = \left[\sqrt{\frac{(\omega^2 LC)^2}{4} + \frac{(\omega RC)^2}{4}} - \frac{\omega^2 LC}{2} \right]^{1/2}$$

¹ Zweig, *op. cit.*

² See, for instance, Everitt and Anner, "Communication Engineering," 3d ed., McGraw-Hill Book Company, Inc., New York, 1956, p. 231.

and the phase shift constant is

$$\beta = \left[\sqrt{\frac{(\omega^2 LC)^2}{4} + \frac{(\omega RC)^2}{4}} + \frac{\omega^2 LC}{2} \right]^{1/2}$$

where ω is the frequency in radians per seconds, while R , L , and C are the resistance, inductance, and capacitance of the line per unit length.

For a hydraulic line in which the flow is assumed to be laminar, the resistance per unit length can be obtained from the Hagen-Poiseuille law [Eq. (10.22)] and becomes

$$R_h = \frac{128\mu}{\pi D^4}$$

The hydraulic inductance per unit length is, by direct analogy with electric circuits, the ratio of pressure drop per unit length to the resulting rate of change of flow. If the pipe has a cross section of A in.² and the fluid has a density of ρ lbs/in.³, then the force exerted on a unit length of fluid by the pressure drop p_l along this unit length is $p_l A$ lbs. Also, the unit mass of the fluid is ρA lbs. Hence by Newton's law

$$p_l A = \frac{\rho A}{g} \cdot \frac{1}{A} \frac{dF}{dt} = \frac{\rho}{g} \frac{dF}{dt}$$

where g is the acceleration of gravity, and F is the flow. Hence,

$$L_h = \frac{p_l}{dF/dt} = \frac{\rho}{Ag}$$

Finally the capacitance per unit length of line is the ratio of a change of oil volume to a change of pressure in the unit length. For a fluid having a bulk modulus of γ psi the change in volume [see Eq. (10.9)] is

$$\Delta q = \frac{A \Delta p}{\gamma}$$

Therefore

$$C_h = \frac{\Delta q}{\Delta p} = \frac{A}{\gamma}$$

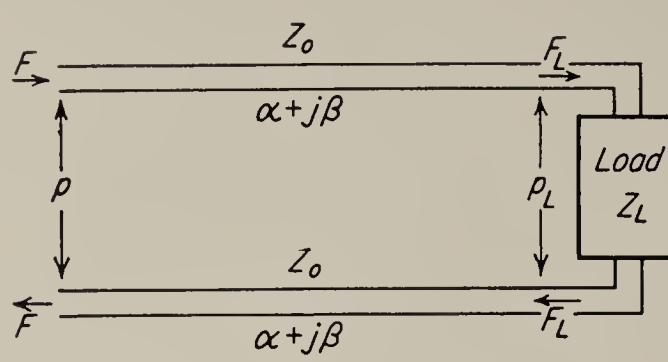
Thus, for the hydraulic transmission line we obtain

$$Z_o = \left[\frac{\left(\frac{128\mu}{\pi D^4} \right)^2 + \left(\frac{\omega \rho}{Ag} \right)^2}{\left(\omega \frac{A}{\gamma} \right)^2} \right]^{1/4} \bigg/ -\frac{1}{2} \tan^{-1} \frac{128\mu Ag}{\pi D^4 \omega \rho} \quad \frac{\text{lbs-sec}}{\text{in.}^5}$$

$$\alpha = \left[\sqrt{\left(\frac{\omega^2 \rho}{2g\gamma} \right)^2 + \left(\frac{64\mu \omega A}{\pi D^4 \gamma} \right)^2} - \frac{\omega^2 \rho}{2g\gamma} \right]^{1/2} \quad \text{nepers/in.}$$

$$\beta = \left[\sqrt{\left(\frac{\omega^2 \rho}{2g\gamma} \right)^2 + \left(\frac{64\mu \omega A}{\pi D^4 \gamma} \right)^2} + \frac{\omega^2 \rho}{2g\gamma} \right]^{1/2} \quad \text{radians/in.}$$

For the special case of the dissipationless line these expressions reduce to



the simpler forms

$$\begin{aligned} Z_o &= \sqrt{\frac{\rho\gamma}{A^2g}} \quad \frac{\text{lbs-sec}}{\text{in.}^5} \\ \alpha &= 0 \\ \beta &= \omega \sqrt{\frac{\rho}{g\gamma}} \quad \text{radians/in.} \end{aligned} \quad (10.24)$$

FIG. 10.16. Hydraulic transmission line.

For a typical hydraulic fluid $\rho = 0.03 \text{ lbs/in.}^3$, $\gamma = 2 \times 10^5 \text{ psi}$ so that Z_o for a standard $7/16$ -in. (ID) copper tube is $26.3 \text{ lbs-sec/in.}^5$ while β is $1.98 \times 10^{-5} \omega \text{ radians/in.}$ The velocity of propagation is ω/β , or approximately $50,000 \text{ in./sec.}$

Zweig¹ has shown that for a symmetrical transmission line such as the one shown in Fig. 10.16 the pressure p across the input terminals is given by

$$p = F_L[Z_L \cosh(\alpha + j\beta)l + 2Z_o \sinh(\alpha + j\beta)l] \quad (10.25)$$

and the input flow is

$$F = F_L[\cosh(\alpha + j\beta)l + (Z_L/2Z_o) \sinh(\alpha + j\beta)l] \quad (10.26)$$

In these equations:

p = pressure difference across the input terminals of the transmission line, psi.

p_L = pressure difference across the load, psi.

F = flow at the input terminals of the line, $\text{in.}^3/\text{sec.}$

F_L = flow at the load, $\text{in.}^3/\text{sec.}$

$Z_L = p_L/q_L$ = hydraulic impedance of the load, lbs-sec/in.^5

l = length in line, in.

Z_o and $\alpha + j\beta$ are the characteristic impedance and propagation constant of one line.

The input impedance of the line is obtained from Eqs. (10.25) and (10.26)

$$Z_i = \frac{p}{F} = 2Z_o \frac{(Z_L/2Z_o) \cosh(\alpha + j\beta)l + \sinh(\alpha + j\beta)l}{\cosh(\alpha + j\beta)l + (Z_L/2Z_o) \sinh(\alpha + j\beta)l} \quad (10.27)$$

It should be borne in mind that a single hydraulic pipe is equivalent to two electric wires, and that therefore the configuration shown in Fig. 10.16 is not exactly equivalent to an electric line connecting a generator to a load. For this reason Eqs. (10.25) to (10.27) are not in exactly the same form as that given in texts on electrical transmission lines. In particular, the factor 2 appearing in several places in the equations arises

¹ Zweig, *op. cit.*

from the fact that there are two lines, each with a characteristic impedance Z_o , and a propagation factor $\alpha + j\beta$.

In practice we are usually interested in obtaining the transfer function of a system including a transmission line. Thus, suppose that the load is a motor of the type discussed in Sec. 10.4, and the source is a variable stroke pump. The first step is to find the hydraulic impedance Z_L presented by the motor. This is done by assuming the motor to have inertia, viscous friction, and possibly leakage. The flow for a given pressure may then be obtained by the methods outlined in Sec. 10.4. The parameters Z_o , α , and β of the hydraulic transmission line are assumed to be known, and thus the input impedance Z_i presented by the combination of motor and transmission line may be computed by Eq. (10.27). This is the load impedance that the pump sees, and therefore, given the characteristics of the pump, F and p can be computed as a function of pump stroke. Finally, Eqs. (10.25) and (10.26) may be used to determine p_L and F_L , and thus the motor output.

When the product $(\alpha + j\beta)l$ is small it is possible to simplify Eqs. (10.25) and (10.26) by using only the first two terms of the Taylor expansion for the hyperbolic functions. The approximate expressions for the equation are, then,

$$p = F_L Z_L \left[1 + \frac{2Z_o}{Z_L} (\alpha + j\beta)l + \frac{1}{2} (\alpha + j\beta)^2 l^2 \right] \quad (10.28)$$

$$\text{and} \quad F = F_L \left[1 + \frac{Z_L}{2Z_o} (\alpha + j\beta)l + \frac{1}{2} (\alpha + j\beta)^2 l^2 \right] \quad (10.29)$$

If the constants are such that the second and third terms in the brackets are small compared to unity, then a solution assuming compressionless, inertialess flow is adequate, and the transmission line has negligible effect on the system. Even if this is not possible, it is usually permissible to neglect the power dissipated in the lines, and to use the expressions for Z_o , α , and β given by Eq. (10.24). This simplification gives

$$p = F_L Z_L \left[1 + j \frac{2\rho}{Z_L A g} \omega l - \frac{\rho}{2g\gamma} \omega^2 l^2 \right] \quad (10.30)$$

$$F = F_L \left[1 + j \frac{Z_L A}{2\gamma} \omega l - \frac{\rho}{2g\gamma} \omega^2 l^2 \right] \quad (10.31)$$

Equations (10.30) and (10.31) indicate that changes in pressure and flow will be chiefly in phase angle if the load is essentially resistive, and will be chiefly in magnitude if the load is reactive.

The expressions obtained here for hydraulic transmission lines may be used also with pneumatic transmission lines, provided that the magnitude of the pressure wave along the lines is small enough so that one can reasonably define a constant bulk modulus. The reader is, however, referred

also to Sec. 12.6 for an empiric relation of a somewhat different form that has been obtained for pneumatic transmission lines.

PROBLEMS

10.1. In a particular pump-controlled hydraulic system like that shown in Fig. 10.1 the pump is rated at 0.738 in.^3 per revolution, 3,600 rpm, and 1,000 psi at full stroke. It may be assumed that the leakage flow is 5 per cent of the load flow at full pressure. The fluid is Univis J43 at 70°F , and the pipe is $\frac{1}{2}$ -in. copper with the motor located 50 ft from the pump. Derive the transfer function from the percentage stroke to motor-shaft-position output for a motor having a displacement of 0.378 in.^3 revolution, a leakage of 5 per cent of full load flow at 1,000 psi, an effective inertia of 50 ft-lb-sec^2 , and a coefficient of viscous damping of 10-ft-lb-sec . Neglect all transmission-line effects except compressibility.

10.2. What is the Reynolds number at full load in Prob. 10.1?

10.3. For a coefficient of energy loss of 0.1 what is the pressure drop in the line in Prob. 10.1?

10.4. Establish the input impedance at the pump of the transmission line and load in Prob. 10.1 as a function of the driving frequency. Assume the line is symmetric.

10.5. Find the transfer function of the system of Prob. 10.1 (a) considering the transmission line effects exactly, (b) using the approximate-transmission-line equations for short lines [Eqs. (10.28) and 10.29)], and (c) using the approximate-transmission-line equations for dissipationless lines [Eqs. (10.30) and (10.31)].

10.6. Figure 10.17 shows a simplified diagram of a common type of hydraulic transmission. The motion of the solenoid armature is amplified by the “stroke servo” consisting of the pilot valve, power piston, and feedback link abc . The length of ab of the feedback link is 0.5 cm, and the length bc is 4.5 cm. The output of the

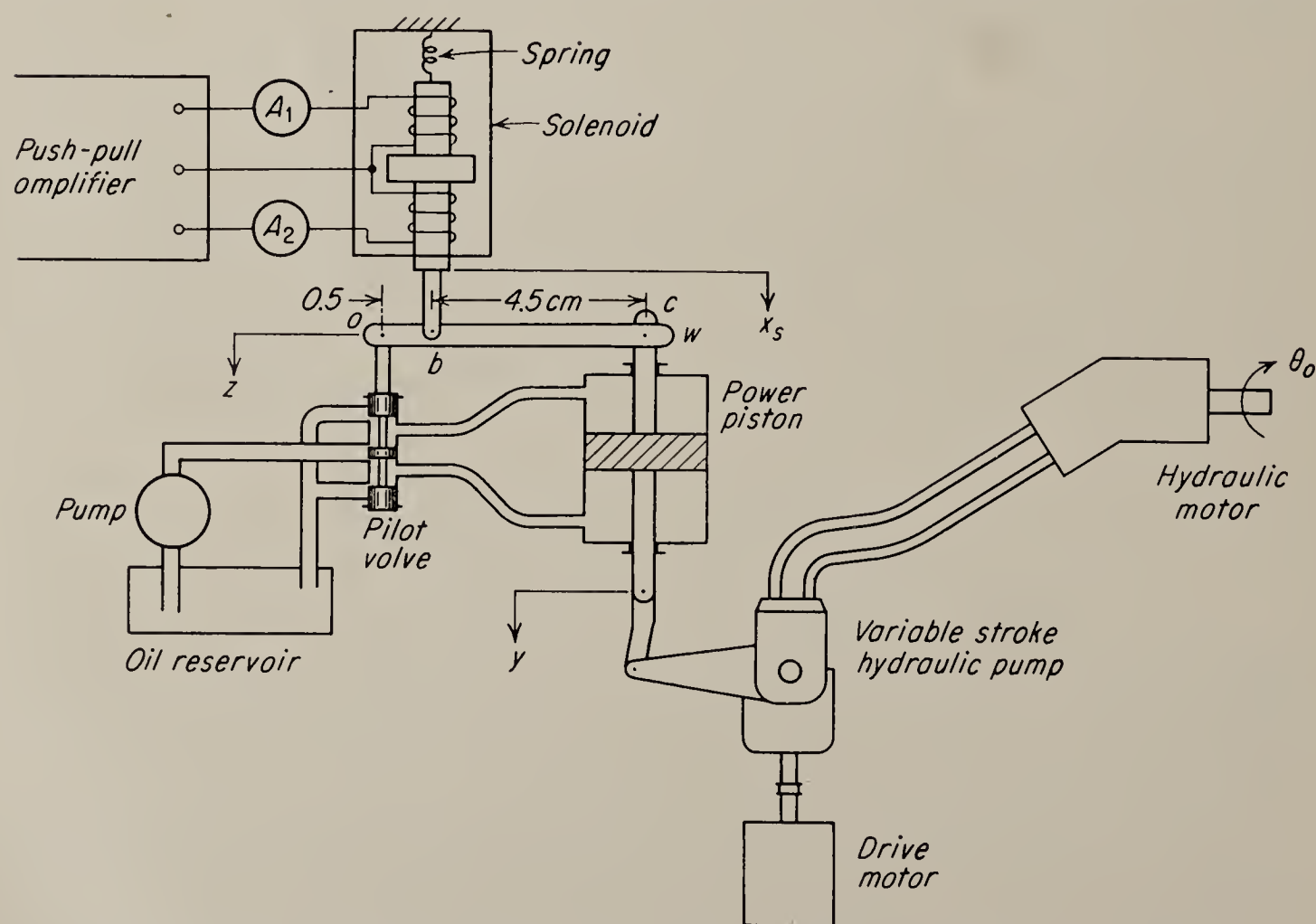


FIG. 10.17. Simplified diagram of a common type of hydraulic transmission.

power piston is applied to the stroke arm of the hydraulic pump and varies the oil delivery to the hydraulic motor. The motor is of the fixed-displacement type; hence its speed is directly proportional to the amount of oil flowing through it.

In order to find the significant data on the stroke servo and solenoid, the linkage is removed from the power piston at point c , and pivot c is then connected to a fixed point so as to act as a fulcrum for the lever abc . Under these conditions it is found that

1. The power piston does not move if the currents in the two halves of the solenoid are equal.

2. The pilot valve moves down 0.002 in. and the piston moves down at the rate of 0.2 in./sec when the currents flowing into the solenoid differ by 1 ma.

3. When the valve is moved down 0.002 in. very suddenly, i.e., if a step function of 0.002 in. is applied to the valve, the initial acceleration of the power piston is 1,000 in./sec².

4. When an alternating current of constant amplitude but varying frequency is forced into the solenoid, a strong resonance is observed in the solenoid response at 40 cps, and at higher frequencies the response drops off as the second power of frequency. At the resonant frequency the amplitude of the x_s swing is three times that observed at very low frequencies.

The following data are available on the pump-and-motor part of the transmission:

1. Pump speed is 4,000 rpm.

2. The pump delivery is proportional to y , being zero when y is neutral (as shown in the figure) and 0.34 in.³ per revolution when y is 0.5 in. from the neutral.

3. The oil displacement of the motor is fixed and equal to 0.38 in.³ per revolution.

4. The motor has an effective moment of inertia of 0.015 lb-in.-sec² and a coefficient of viscous friction of 0.1 lb-in.-sec.

5. Owing to oil leakage from the high- to the low-pressure lines, the motor speed will decrease by 3 rpm per inch-pound of torque applied to the output shaft.

6. Total oil volume in both oil lines is 12 in.³.

7. The effective compressibility of the oil used (this includes the stretch in the tube walls) is 4.4×10^{-6} in.²/lb.

Assume linearity throughout the analysis of this system. Also it is permissible to assume that the reaction force of the pump on the power piston and the hydraulic reaction force of the pilot valve are negligible. Further, the volume of oil in the stroke servo is so small that compressibility effects in that part of the transmission may be neglected.

Problem: Find the complete expression for the transfer function $\hat{\theta}/\hat{i}_s$, where θ is the position of the output shaft of the hydraulic motor, and i_s is the difference in currents in the two halves of the solenoid. Sketch the asymptotic and exact frequency-response curve, and the phase-shift curve for the transmission. Neglect effects above frequencies of 1,000 radians/sec.

CHAPTER 11

VALVE-CONTROLLED HYDRAULIC SYSTEMS

11.1. Introduction. The use of a variable-stroke pump to control a hydraulic motor as discussed in Chap. 10 is unsatisfactory if there are several hydraulic actuators that are supplied from the same pump and that must be independently controlled. Usually there is one hydraulic power supply that operates at constant pressure in this type of system,

and the actuators are individually controlled by pilot valves.

In this chapter we shall consider the valve-controlled system and the components that are unique to it. The motors and pumps have been considered in Chap. 10. Control valves and pressure regulators will be discussed here. As in Chap. 10, the devices will be introduced by a discussion of a simplified system.

In Fig. 11.1 is shown a hydraulic actuator or stroke amplifier such as might be used to operate the variable-stroke pump in Sec. 10.3. The valve is a standard three-land spool type and is actuated from the input through the linkage with a pivot at point W . Mathematically Y and W are the same point and are named separately only for clarity. If the

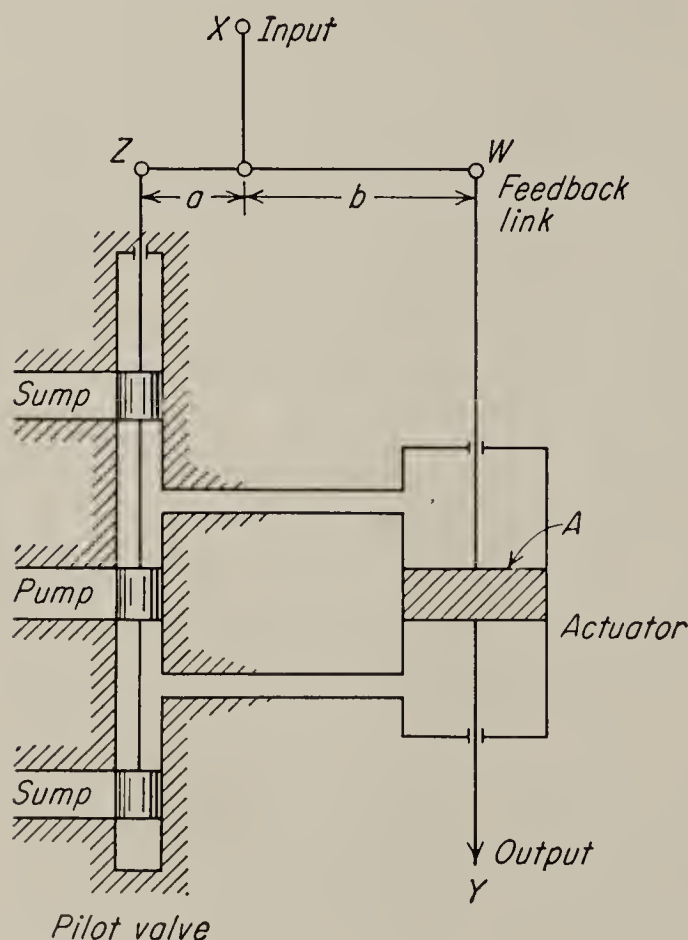


FIG. 11.1. A simple hydraulic valve and actuator.

hydraulic fluid is considered incompressible, point W cannot move with the valve closed; thus W may be considered a pivot. If X moves down, Z moves down, and high-pressure fluid is ported to the top of the actuator. At the same time the bottom of the actuator is opened to sump. The actuator moves down, following the input. As W moves down, the linkage moves about X as a pivot, carrying Z up and closing the valve. Let us assume that the valve flow is proportional to valve opening Z ; then

$$Z = \frac{a+b}{b} X \quad \text{in.} \quad (11.1)$$

$$\text{and} \quad F = K_v Z \quad \text{in.}^3/\text{sec} \quad (11.2)$$

where K_v is the constant relating motion to flow and has the dimensions of (in.³/sec)/in. Flow may be converted to velocity of the actuator by dividing by the area A of the piston. Thus the transfer function from valve position Z to Y is

$$\frac{\hat{Y}}{\hat{Z}} = \frac{K_v}{As} \quad (11.3)$$

FIG. 11.2. Block diagram for the hydraulic actuator shown in Fig. 11.1.

The transfer function through the feedback link is the lever ratio

$$\frac{\hat{Z}}{\hat{W}} = -\frac{a}{b} \quad (11.4)$$

where the minus sign shows the change in direction. Figure 11.2 shows a block diagram for the complete unit. The closed-loop transfer function from X to Y is thus

$$\frac{\hat{Y}}{\hat{X}} = \frac{a+b}{b} \frac{K_v/As}{1 + (aK_v/bAs)} = \frac{(a+b)/a}{(bA/aK_v)s + 1} \quad (11.5)$$

The valve in this example primarily controls the hydraulic flow. It is also possible to design a valve that primarily controls pressure. Figure 11.30 shows an example of a pressure-control valve of the flapper type discussed in Sec. 11.9. The position of the flapper controls p_2 , the chamber pressure which accelerates the actuator if it is unloaded. The transfer function from flapper position to unloaded actuator position thus contains two integrations.

11.2. Spool-type Pilot Valves. The spool valve shown in Fig. 11.1 is one of several spool-type valves that are classified as four-way valves. In the four-way valve there are four orifices that are critically spaced with respect to one another. In Fig. 11.3 are shown two more four-way valves. It will be noted that these valves involve no basic modification in operation nor simplification of manufacturing. The number of critical dimensions remains the same as with the valve in Fig. 11.1.

A somewhat different type of operation is involved in the two-way spool valves shown in Fig. 11.4. Here there are only two orifices, and the valves are restricted to operation with unequal-area actuators. The high-pressure supply is connected to the smaller-area side of the actuator, and the valve controls the flow and pressure to the larger-area side. The

derivation of the transfer function from valve position to actuator position is left as an exercise for the reader.

The clearances and tolerances in pilot valves are held to as close as 0.0001 in. to minimize leakage. Proper fitting of the valve spool is a

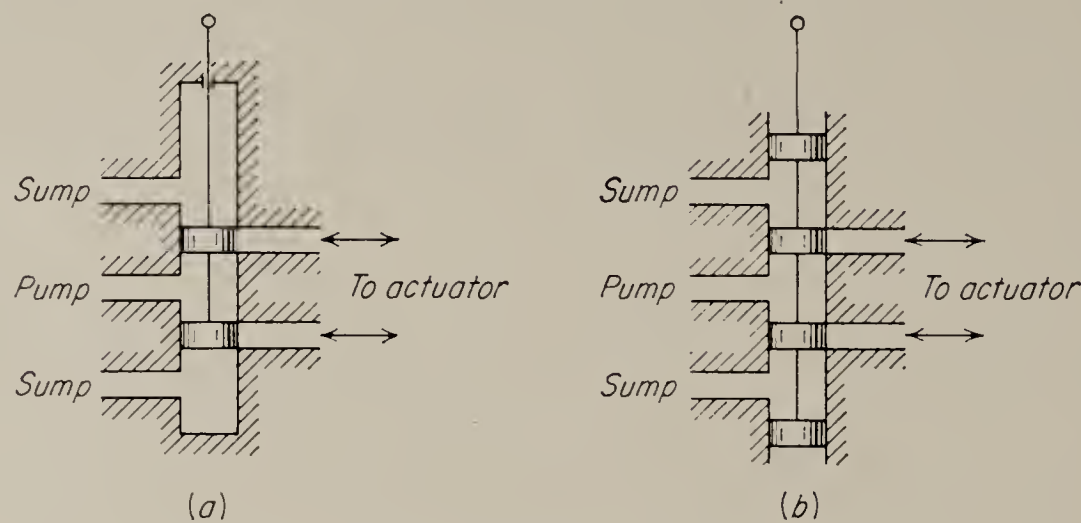


FIG. 11.3. Two types of four-way spool valves.

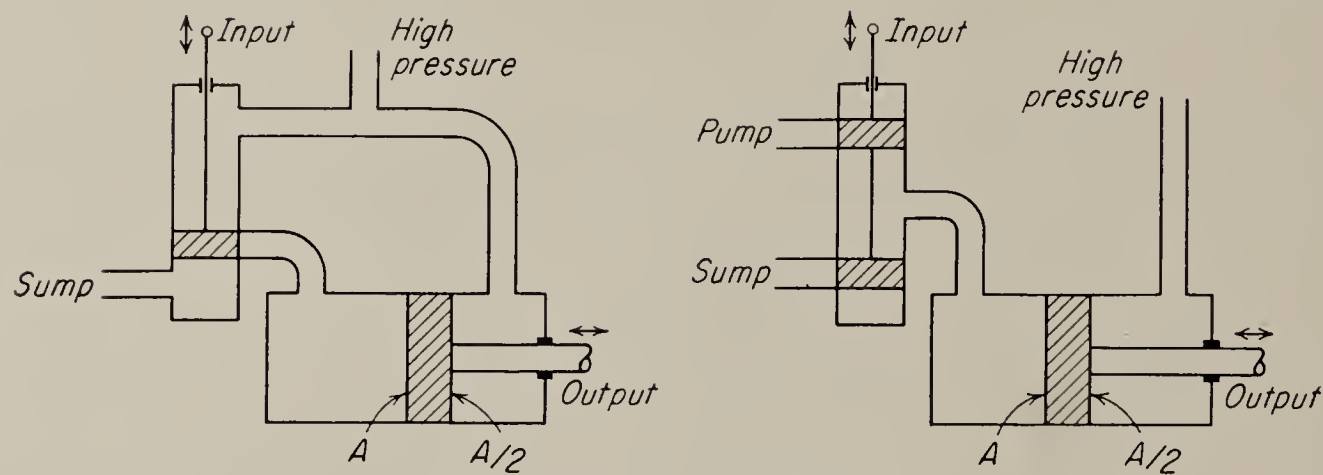


FIG. 11.4. Two-way spool valve.

difficult manufacturing problem. Exact alignment of the valve lands and ports is also a major problem. A valve in which the edges of the lands exactly meet the edges of the ports is called a *zero-lap valve*. It is impossible to manufacture a valve with exactly zero lap, and if the valve

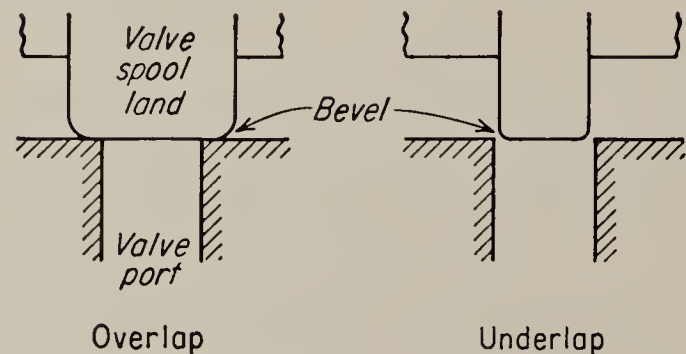


FIG. 11.5. Overlap and underlap on lands of valve spool.

is to be constructed at all, the designer must allow at least a certain minimum underlap or overlap (see Fig. 11.5). Even if the valve spools could be manufactured with zero lap and no bevel, the initial knife-edge would rapidly round off in use.

A valve with overlap has a *dead zone* equal to the amount of overlap.

This nonlinearity results in loss in sensitivity near the center and may cause instability in control systems. Usually, therefore, valves with a small amount of underlap are preferred for control-system applications. The disadvantage of the underlapped valve is the leakage that exists even when the valve is centered.

In the manufacture of a four-way valve spool, there are three critical dimensions that must be controlled, without including the diameter of the lands. These are the distances between the four orifices. These critical dimensions cannot be adjusted independently. One solution to this difficult manufacturing problem is to build a valve with two separate spools and to connect the two spools together by an adjustable linkage. While eliminating the interdependence between dimensions, this method introduces several other problems. The resulting valve is bulkier, and care must be exercised if the play in the link is to be smaller than the manufacturing inaccuracies of the original valve.

Although round valve ports are easier to manufacture than rectangular ports, the latter are preferred in high-performance systems because the variation of port area is then a linear function of stroke.

Several proposals have been advanced that avoid, to some extent at least, the extremely close tolerances required by high-gain spool-type valves. For instance, valves that operate on entirely different principles, such as slide valves and nozzle valves, are used. The tolerances of slide valves and nozzle valves, as discussed below, are not quite so critical as those of spool valves. The nozzle valve (Sec. 11.10), however, is characterized by relatively high leakage, and the decrease in manufacturing costs of the slide valve (Sec. 11.11) is largely canceled by the relatively expensive linear torque motor required.

11.3. Pulsed Operation of Hydraulic Valves. A possibility for the reduction of required manufacturing accuracy in spool valves lies in a very interesting mode of operation proposed by Jackson¹ and Chubbuck.²

The input signal to the valve is in the form of a pulse-length-modulated wave rather than a continuous d-c value. That is, the input is a train of pulses, each of constant amplitude and large enough to cause the valve spool to move from full open in one direction to full open in the other. The waveshapes are shown in Fig. 11.6. In Fig. 11.6a is shown the ideal characteristic of the conventional proportional valve; this ideal is only approximated in practice. In Fig. 11.6b is shown the input to the valve for no signal in the pulse operation mode. The signal is converted into the form of *pulse-length modulation* (PLM) by a modulator preceding the valve. In Fig. 11.6c is shown the pulse wave train for a signal of maximum amplitude applied to the PLM valve driver. The advantages of the PLM mode appear to be several:³

1. The torque motor or electric actuator of the valve can be optimized

¹ K. R. Jackson, U.S. Patent No. 2,655,940, 1950.

² J. G. Chubbuck, "Acceleration Switching Hydraulic Servo," Johns Hopkins Univ. Appl. Phys. Lab. Tech. Rept., May, 1955, and *Control Eng.*, vol. 4, no. 3, March, 1957.

³ J. E. Gibson, F. B. Tuteur, and T. Mapes, A New Hydraulic Servo Valve for Pulse-length-modulation Operation, ASME paper no. 57-A-128.

for maximum force per ampere. This is not so in the conventional valve, since efficiency must be sacrificed in order to obtain linearity. Thus in the PLM mode the actuator may be made smaller and lighter.

2. In conventional valves, the flow gain is usually reduced at small signal amplitudes. This does not occur in PLM mode operation.

3. Small-signal sticking is eliminated since in a sense the PLM mode provides the spool with a massive *dither* (see Sec. 11.6).

4. Small-opening Bernoulli forces, discussed in Sec. 11.5, become unimportant.

5. Valve lock, discussed in Sec. 11.6, is a less serious problem because the valve is kept in motion and because it is not necessary to design for exceedingly small tolerances, as is required in the linear mode. Furthermore, the emphasis in the design of the torque motor is on maximum force rather than on linearity. The increased force tends to override the normal small-opening effects.

6. Manufacturing tolerances may be relaxed, since the flow need not be linear with valve stroke. Round valve ports, for instance, are perfectly satisfactory.

7. Since the torque motor for the PLM mode can be designed efficiently, a single-stage valve becomes a possibility in many applications where two-stage valves were formerly

required. The single-stage PLM valve with its PLM modulator will usually be more economical than the precision two-stage linear valve and its driver.

Naturally the pulsed valve has several disadvantages compared to linear valve operation:

1. The output flow is in the form of pulses of fluid; thus the load must be such that its low-pass filtering effect is sufficient to smooth the flow adequately. Fortunately, many loads do provide adequate smoothing. One case where pulsed operation would not be desirable would be where the load was driven, through a multimesh gear train, by a rather light hydraulic motor. The motor could follow the high-frequency pulsing of

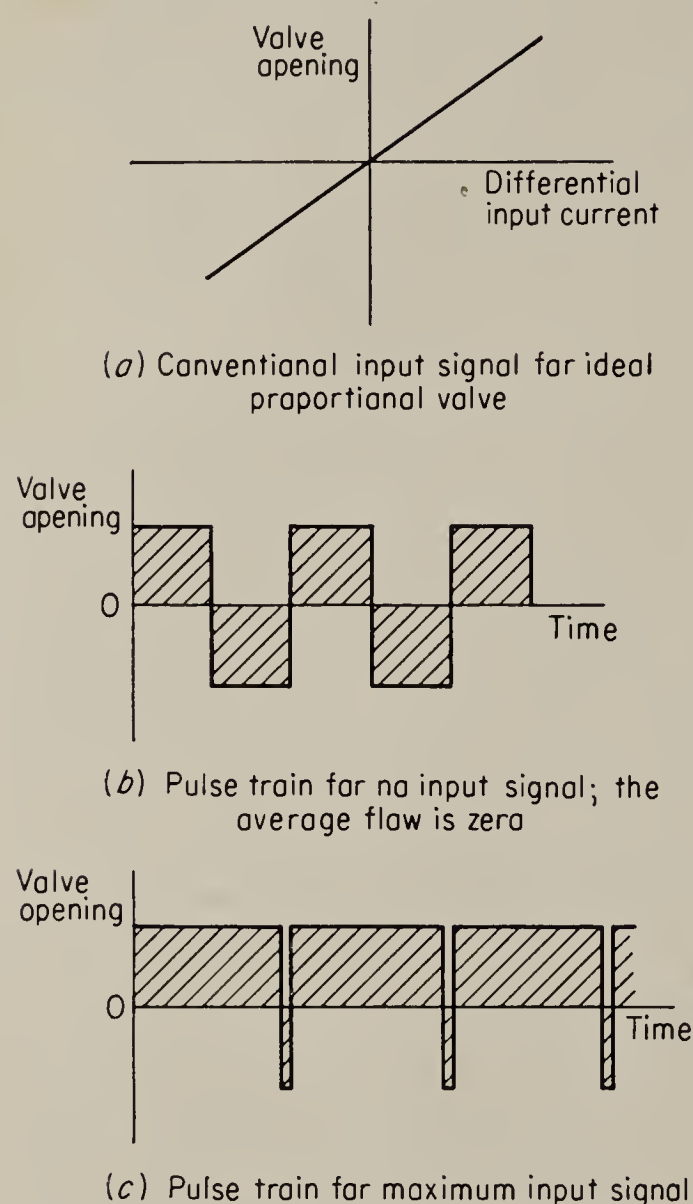


FIG. 11.6. Input waveshapes for pulse operation of hydraulic valves.

the fluid and destroy the gear train by vibrating within its initial backlash limits.

2. An original disadvantage of the PLM mode was the rather complex electronic pulse-length modulator required. When it is realized, however, that the magnetic amplifier is almost ideally suited to the PLM mode, this problem is eliminated.¹ The normal output waveshapes of magnetic amplifiers are portions of the sine wave of the a-c supply frequency. Thus the magnetic amplifier may drive the valve directly by controlling the instant of conduction.

Tests have shown that pulse operation will also improve the performance of high-precision valves originally designed for linear operation. The frequency response will usually be improved, and resonances and dips due to deadband and friction will be eliminated.

The pulsed mode of operation is by no means limited to spool valves. It may be applied to all the common designs. Indeed, the concept of pulse operation opens up possibilities for entirely new types of valves and hydraulic systems.

In a sense, the pulse mode of operation removes the burden of high precision from the mechanical design of the valves and increases the requirements on the electric drive for the valve. Fortunately, the problems involved in designing electronic circuits to convert a continuous signal into the form of a pulse-length-modulated wave are relatively easy to solve. The circuit can be a modified form of the Eccles-Jordan trigger circuit and can be built with six or seven transistors including the power stage, or a compact magnetic amplifier can be designed to provide the required drive.

The basic pulse repetition rate must be at least twice² the highest frequency that the valve must follow, and for negligible phase shift due to the pulsing, the pulse repetition rate should be about ten times the highest signal frequency. In practice, a factor of three to five times the highest signal frequency is satisfactory. If the repetition rate is too high, the valve will be unable to follow it, and if the repetition rate is too low, the frequency response of the valve is limited.

A single-stage servo valve designed for PLM operation with a 400-cps repetition rate provided a frequency response that was flat to above 100 cps.³ The repetition rate of 400 cps was chosen so that the magnetic-amplifier driver could operate from a 400-cps supply. The maximum theoretical operating frequency would be one-half this value, or 200 cps. Actual operation was successful up to about 150 cps.

¹ *Ibid.*

² B. M. Oliver, J. R. Pierce, and C. E. Shannon, The Philosophy of PCM, *Proc. IRE*, vol. 36, no. 11, pp. 1324-1331, 1948.

³ Gibson, Tuteur, and Mapes, *op. cit.*

As was noted above, the output flow of the PLM-operated valve must be smoothed by the load. In this sense, the PLM mode of operation may be considered to be the hydraulic equivalent of the relay amplifier (see Chap. 3) operated in the stable oscillation mode. In fact, the static characteristic found experimentally by Chubbuck for the pulsed valve is identical in form to the static characteristic of the relay amplifier derived in Chap. 3 and shown in Fig. 3.42.

Indeed analogues of on-off electronic circuits and devices will provide the hydraulic-servo designer with many interesting possibilities for investigation.

11.4. Flow-Pressure Relations in Orifices and Valves. In Sec. 11.1 it was assumed that the flow through the pilot valve was directly proportional to valve opening. Actually this is true only if the pressure drop across the valve is ignored or assumed constant. This assumption is not supported by the actual physical facts. Let us consider the actual relationships. After defining certain general flow relations in orifices we shall discuss more specific relations in spool-type valves.

By definition, a valve or *orifice* is a portion of the flow path that has a restricted flow area and across which a large pressure drop may be developed. It may be shown by the law of conservation of energy¹ that theoretical volumetric flow through an orifice is

$$F = A \sqrt{\frac{2gp}{\rho}} \quad (11.6)$$

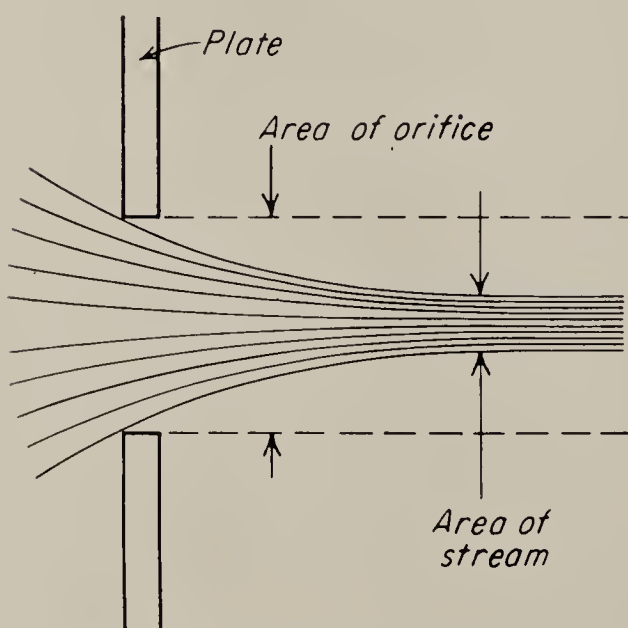


FIG. 11.7. Orifice in a thin plate, showing the contraction effect.

where p is the pressure drop across the orifice, A is the actual cross section of the orifice, g is the acceleration of gravity, and ρ is the density of the fluid.

Actually, the flow from an orifice is less than the theoretical flow for two reasons. First, owing to friction in the flow of the fluid through the orifice, the flow is reduced by a factor C_v , the *velocity coefficient*. Second, the stream is contracted as it moves through the

orifice because most of the particles of fluid are moving in a curved path, as shown in Fig. 11.7. The effective minimum cross section of the stream is therefore less than the orifice cross section. The ratio of effective area to orifice area is called C_A , the *contraction coefficient*. The actual flow through an orifice is thus

¹ Dodge and Thompson, "Fluid Mechanics," McGraw-Hill Book Company, Inc., New York, 1937.

$$F = C_v C_A A \sqrt{\frac{2gp}{\rho}} = CA \sqrt{\frac{2gp}{\rho}} \quad (11.7)$$

where C is the *discharge coefficient* of the orifice. The value of C depends on the shape of the orifice as well as on the area of the orifice and the operating pressure. For usual operating conditions, pilot valves in control systems have been found to have discharge coefficients between 0.6 and 0.8. The higher value applies for orifices with rounded edges, a shape which reduces separation and turbulence in the flow.

Flow separation occurs at sharp bends or edges around which the fluid must pass. *Flow separation* is the curling back or eddying that occurs following such bends or edges; Fig. 11.8 shows several examples of flow

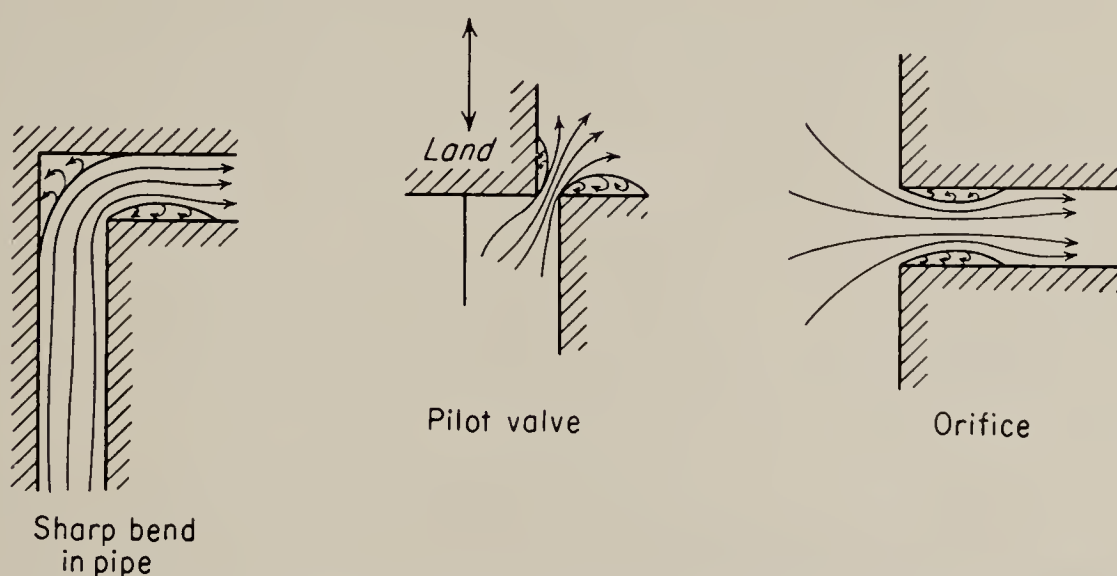


FIG. 11.8. Examples of flow separation in typical control-system applications.

separation. Flow separation causes abnormally high pressure drops and energy losses in the system. Smoothly rounded corners will decrease or eliminate separation. In most control valves it is necessary to tolerate separation, since the edges must be sharp.

Cavitation, another hydraulic phenomenon similar in certain of its effects to separation, occurs in a hydraulic device whenever the fluid pressure is reduced to the vapor pressure of the fluid. The fluid then passes into its vapor state. The small vapor cavities or bubbles are carried into regions of higher fluid pressure, where they collapse. Cavitation causes a higher pressure drop than would otherwise be expected and resultant energy losses. What may be equally important, the collapse of the cavities is accompanied by sudden shock forces that may reach high magnitude. These shock forces may induce metal fatigue and pitting of the surfaces exposed to the cavitation. There also can be an erosion effect from bubbles forming in the pores of the metal and bursting when the metal is carried into a high-pressure region.

Cavitation is occasionally useful in hydraulic-system design. Certain devices must be protected by limiting the maximum flow, should the

in the relation. However, if the port is circular, the area as a function of stroke is as shown in Fig. 11.10. Even in this case the cross-sectional area is approximately proportional to stroke.

Let us now consider the effect of four orifices in a bridge-type circuit of the sort found in the typical spool-type valve (see Fig. 11.11). We shall

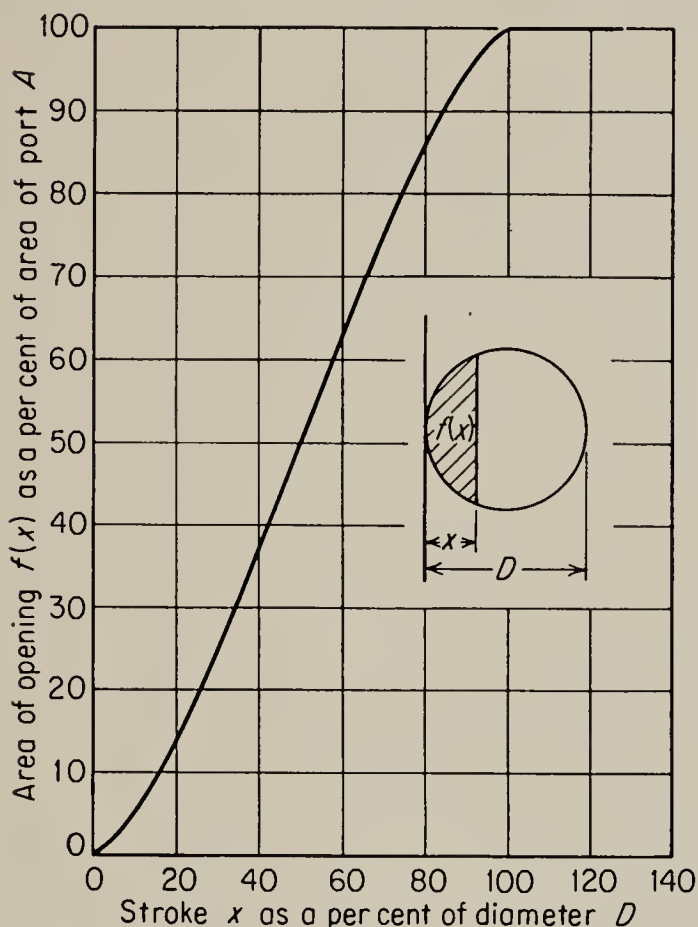


FIG. 11.10. Cross-sectional area of a circular valve port as a function of stroke, normalized to total area and full stroke.

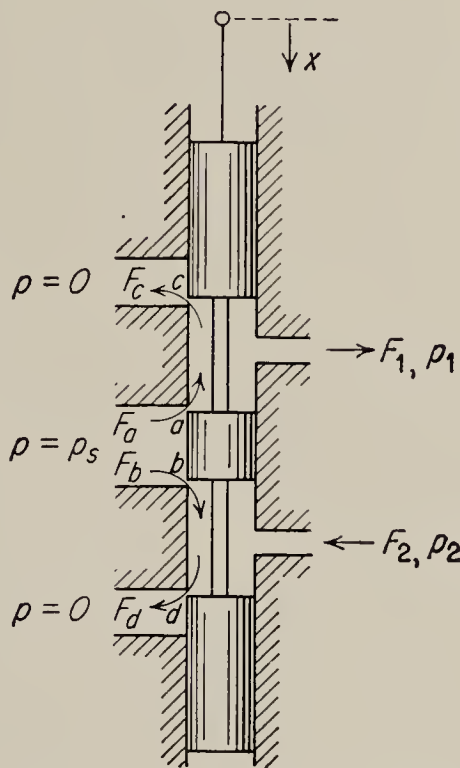


FIG. 11.11. A typical spool-type valve with underlap x_0 .

assume that the orifice area is proportional to the displacement x , and that the valve is symmetric and has an underlap x_0 . Also, it will be assumed that the valve operates with a constant supply pressure p_s and that the pressure in the return line is zero. From Eq. (11.7) we may write for the four orifices

$$\begin{aligned}
 F_a &= k(x_0 + x) \sqrt{p_s - p_1} & x_0 + x > 0; \text{ otherwise } F_a &= 0 \\
 F_b &= k(x_0 - x) \sqrt{p_s - p_2} & x_0 - x > 0; \text{ otherwise } F_b &= 0 \\
 F_c &= k(x_0 - x) \sqrt{p_1} & x_0 - x > 0; \text{ otherwise } F_c &= 0 \\
 F_d &= k(x_0 + x) \sqrt{p_2} & x_0 + x > 0; \text{ otherwise } F_d &= 0
 \end{aligned} \tag{11.8}$$

and the load flow will be

$$F_L = F_a - F_c \tag{11.9}$$

Thus we may write, combining Eqs. (11.8) and (11.9), that

$$F_L = k(x_0 + x) \sqrt{p_s - p_1} - k(x_0 - x) \sqrt{p_1} \tag{11.10}$$

Equation (11.10) can be simplified if we assume that the valve is symmetrical. This implies that $p_s = p_1 + p_2$. Also, in general, the load

pressure $p_L = p_1 - p_2$. Equation (11.10) may therefore be rewritten as

$$F_L = k(x_o + x) \sqrt{\frac{p_s - p_L}{2}} - k(x_o - x) \sqrt{\frac{p_s + p_L}{2}} \quad (11.11)$$

or if we make the natural definition $p_v = p_s - p_L$, where p_v is the valve pressure, then Eq. (11.11) may be written

$$F_L = k(x_o + x) \sqrt{\frac{p_v}{2}} - k(x_o - x) \sqrt{p_s - \frac{p_v}{2}} \quad (11.12)$$

Equation (11.12) holds in the underlap region, where $|x| < x_o$. Outside of the underlap region one term of the equation falls out, and only the first or second term remains, depending on whether $x > x_o$ or $x < -x_o$.

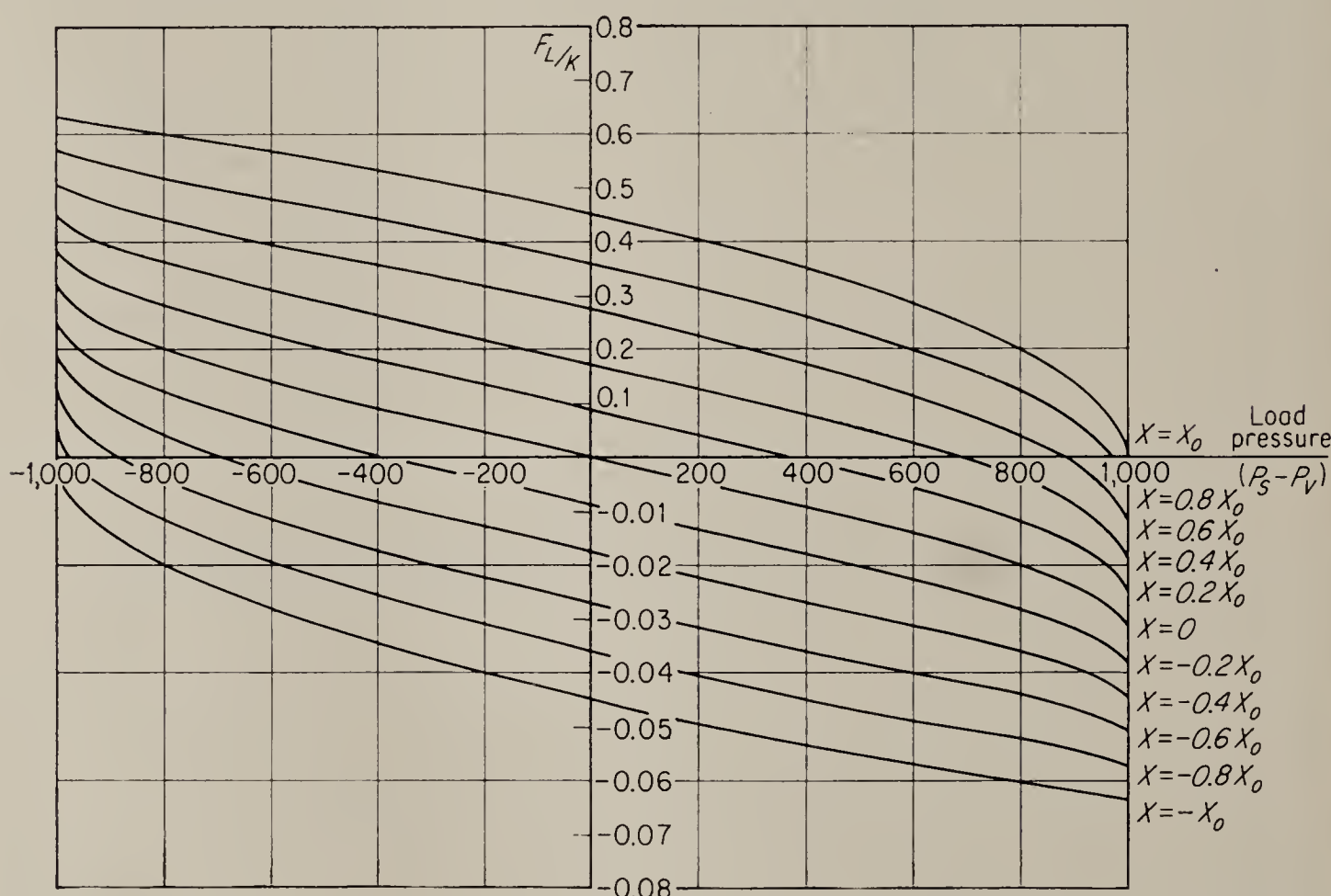


FIG. 11.12. Flow plotted against pressure for an underlapped valve operating in the underlapped region and with constant supply pressure.

The reason for manipulating the relations into the form of Eq. (11.12) is that characteristic curves for the valve may be calculated and employed in an analysis, as shown in Sec. 11.7. In Fig. 11.12 is shown the characteristic calculated from Eq. (11.12) of an underlapped valve operating in the underlapped region with a constant supply pressure. It is assumed here that area of the valve orifices is proportional to stroke. The orifice coefficient used was 0.7, which is a reasonable assumption. It will be noted that the characteristic is essentially linear in this region. This is one of the reasons that the underlapped valve is preferred to the overlapped valve.

In Fig. 11.13 is shown the characteristic of the valve outside of the underlapped region. In this case one of the terms of Eq. (11.12) was dropped, as mentioned above, in order to calculate the characteristic. The characteristics for positive and negative stroke are superimposed as shown, for convenience in graphical analysis. Figure 11.13 may also be used for a zero-lapped valve directly and for an overlapped valve if the portion of the stroke in the overlap region is accounted for separately. The calculation of valve characteristics for a constant supply flow rather than constant supply pressure is left as an exercise for the reader.

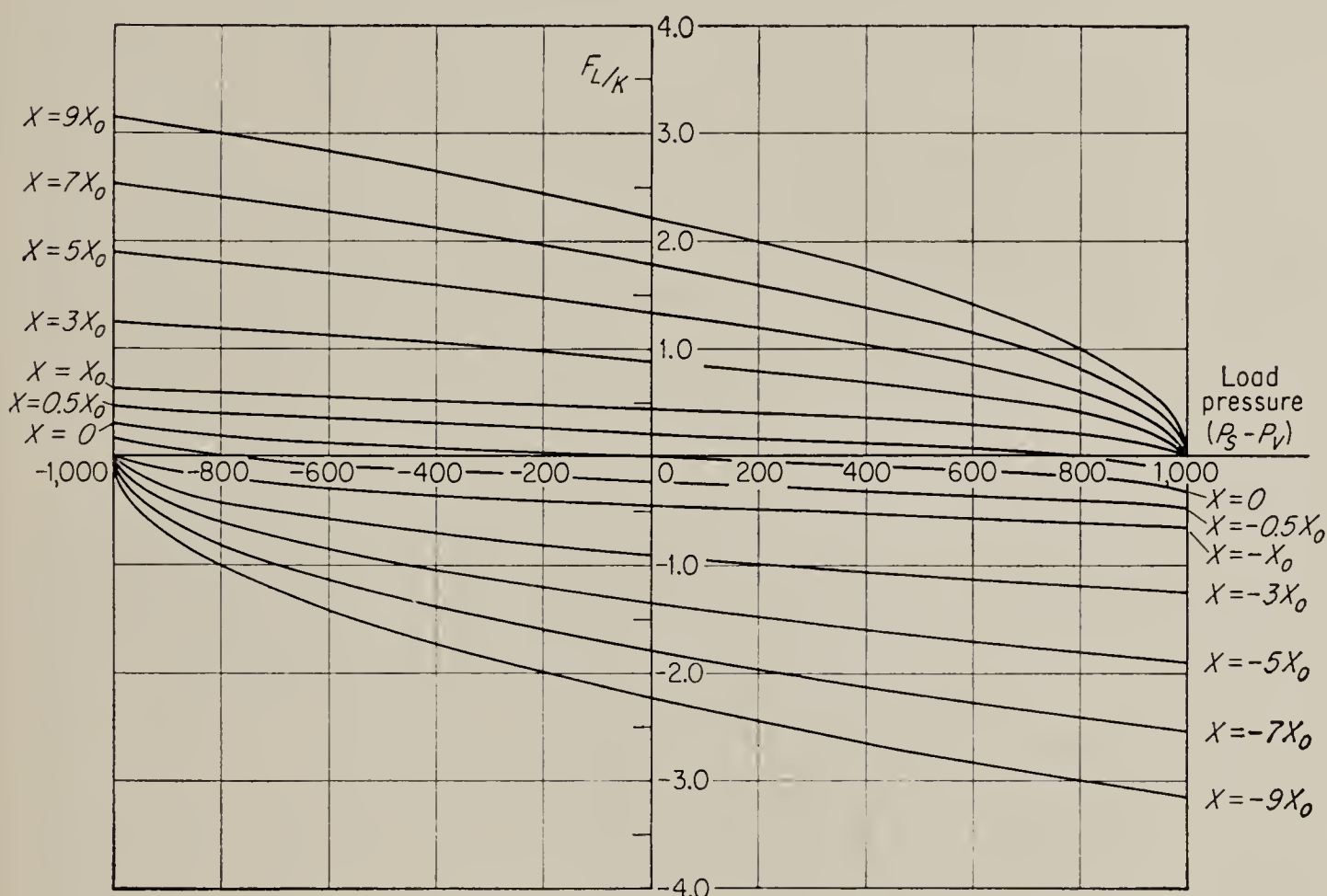


FIG. 11.13. The flow-pressure characteristic of an underlapped, spool-type valve for operation beyond the underlap region.

11.5. Axial Hydraulic Reaction Forces in Spool-type Valves. The flow of fluid through the orifices of a control valve causes a hydraulic reaction force, also referred to as Bernoulli force. This force has two components, a radial and an axial component. The *radial component* tends to push the valve spool sideways against the sleeve and causes sticking. It is usually minimized by locating the valve ports symmetrically about the spool, but a component may remain if the leakage paths between spool and sleeve are not symmetrical. The radial component of force is considered in more detail in Sec. 11.6.

The *axial force* is usually in such a direction as to close the valve. This force has been considered in some detail by Lee and Blackburn.¹ The

¹S. Y. Lee and J. F. Blackburn, Contributions to Hydraulic Control I: Steady-state Axial Forces on Control-valve Pistons, *Trans. ASME*, vol. 74, p. 1005, 1952.

valve configuration studied is shown in Fig. 11.14. The flow into the chamber through de is essentially unrestricted if the opening x is small compared to the circumferential length of the orifice; and if the flow is assumed to be irrotational, nonviscous, and incompressible, the solution of the flow pattern in the chamber is the solution of Laplace's equation for the configuration. This solution¹ yields a θ angle of 69° when the valve is square ($\phi = 90^\circ$) and when there is no radial clearance. With θ known, the force on the valve spool can be found by conservation of momentum of the flow in the valve chamber. Since the area de is large with respect to the orifice area, the velocity of flow through de is negligible as compared with the velocity of flow through the orifice. Axial pressures on the ends of the chamber are equal. Thus the flow through the orifice produces a change in momentum that is balanced only by an axial piston force. This change of momentum may be expressed in terms of the fluid flow by

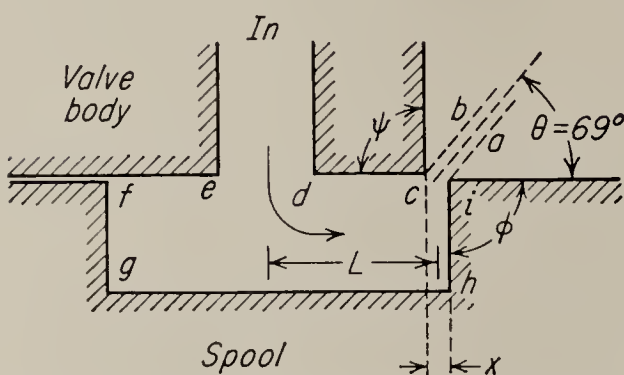


FIG. 11.14. Configuration for derivation of axial reaction force.

where Mv is the momentum, F the flow rate, ρ the density, v the velocity of the fluid at the *vena contracta* (area ab in Fig. 11.14), and g the acceleration of gravity. Note that F is considered positive when the flow is out of the controlled orifice. The axial component of the force is therefore given by

$$\mathcal{F}_{\text{axial}} = - \frac{F\rho v}{g} \cos \theta \quad (11.13)$$

The minus sign arises from the fact that the expression is for the force by the fluid on the container; this is the negative of the force required to accelerate the fluid. By Bernoulli's equation the velocity v is given by

$$v = \sqrt{\frac{2gp}{\rho}}$$

where p is the pressure drop across the orifice. Also, the flow rate F may be expressed in terms of p by Eq. (11.7). Hence the axial force becomes

$$\begin{aligned} \mathcal{F}_{\text{axial}} &= -CA \sqrt{\frac{2gp}{\rho}} \cdot \frac{\rho}{g} \cdot \sqrt{\frac{2gp}{\rho}} \cdot \cos \theta \\ &= -2CAp \cos \theta \end{aligned} \quad (11.14)$$

¹ R. Von Mises, "Berechnung von Ausfluss- und Überfallzahlen," *Z. Ver. deut. Ing.*, vol. 61, p. 494, 1917.

where C is the orifice coefficient and A is the cross-sectional area of the orifice. For an orifice coefficient of 0.6, the axial force per orifice is

$$\mathcal{F}_{\text{axial}} = 0.43 Ap \text{ lbs}$$

if the pressure is expressed in psi and the area in square inches. In a symmetrical valve the inlet and outlet orifices are identical and are in series. Therefore, the total force is twice the force per orifice. However, the total pressure drop across the valve is also twice the pressure per orifice. Thus, the total axial force may be written

$$\mathcal{F}_T = 0.43 p_v$$

where p_v is now the total valve-pressure drop.

Note that in Fig. 11.14 both F and $v \cos \theta$ are positive; hence the force is negative (to the left) and tends to close the valve. If the flow is reversed, both F and $v \cos \theta$ would be negative, and the force would still be in the same direction.

Lee and Blackburn¹ have investigated the more realistic case of a valve with radial clearance, and with land edges of finite radius. Both of these imperfections of the valve result in an increased orifice area and a decreased angle θ of the jet. Hence they both tend to increase the force generated per unit displacement of the valve spool, particularly for small spool displacements. A typical set of curves of force versus displacement for different clearances and radii is shown in Fig. 11.15. The pressure drop across the orifice is constant in this figure.

The hydraulic reaction force given in Eq. (11.14) is that obtained for steady flow. When the flow is changed by the opening or closing of the valve, an additional force is generated. It is not unreasonable that such a force should exist. Thus, consider the valve configuration shown in Fig. 11.14 and assume that the valve spool is suddenly moved slightly to the left so as to reduce the orifice opening. The average velocity of the oil volume in the chamber must then be reduced, and the resulting momentum change would be expected to resist the leftward motion of the valve. The same result is obtained if the valve spool is moved to the right: the momentum change of the fluid will resist the spool motion. On the other hand, suppose that the fluid flow is reversed: into the chamber at ab and out at ed . Again assume that the spool moves to the left. Now the momentum of the fluid is in the opposite direction, and tends to pull the spool along with it. Hence the transient force is in a direction to aid the spool motion. Lee and Blackburn² have shown that the total

¹ Lee and Blackburn, *loc. cit.*

² S. Y. Lee and J. F. Blackburn, Contributions to Hydraulic Control II: Transient-flow Forces and Valve Instability, *Trans. ASME*, vol. 74, pp. 1013–1016, 1952.

reaction force at each orifice may be written

$$\mathcal{F}_{\text{axial}} = -\frac{\rho}{g} F v \cos \theta + \frac{\rho}{g} L \frac{dF}{dt} \quad (11.15)$$

where L is the axial distance between centers of incoming and outgoing flows (see Fig. 11.14), and where all other quantities are as defined previously. The distance L is negative when the flow leaves the chamber through the controlled orifice (as in Fig. 11.14), and it is positive when the flow enters through the controlled orifice. The flow rate F is again assumed to be positive for flow out of the chamber. The first term of Eq. (11.15) is identical with Eq. (11.13) and represents the steady-state

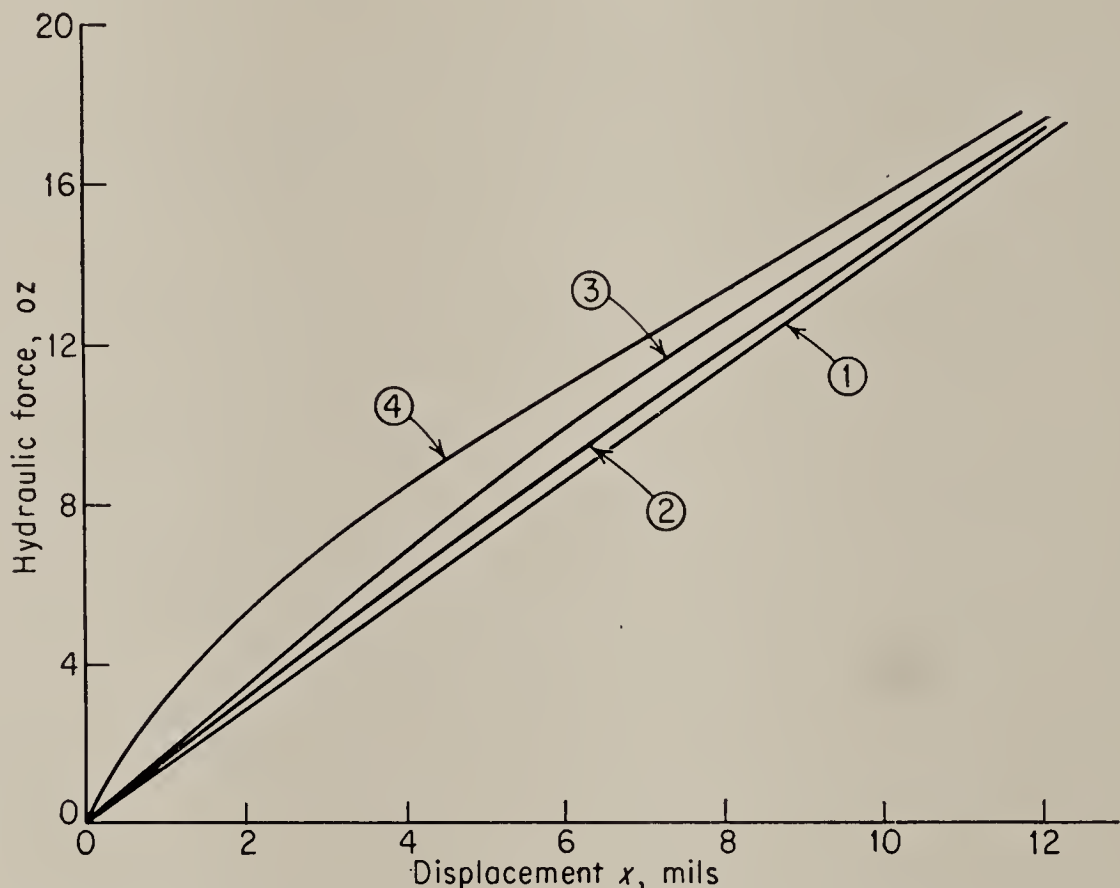


FIG. 11.15. Hydraulic-reaction force curves: (1) Ideal case, clearance = 0, radius = 0; (2) clearance = 0.0001 in., radius = 0; (3) clearance = 0.0005 in., radius = 0.0003 in.; (4) clearance = 0.0003 in., radius = 0. (From Lee, Blackburn)

component of the force. The second term is the transient force. For the situation shown in Fig. 11.14 the transient force is negative, i.e., in a direction to close the valve, when the valve is being opened and the flow is increased. This is in accordance with the qualitative discussion of the transient-force phenomena given above.

It is again convenient to convert Eq. (11.15) to a form such that the force is a function of pressure, rather than flow. Using Eq. (11.7) we obtain

$$\mathcal{F}_{\text{axial}} = -2CAp \cos \theta + CL \sqrt{\frac{2\rho}{g} p} \frac{dA}{dt} + CLA \sqrt{\frac{\rho}{2gp}} \frac{dp}{dt}$$

If, as in Sec. 11.4, we make the further assumption that A is proportional

to x , we can set $A = wx$ where w is the effective width of the orifice. Then the axial force becomes

$$\mathcal{F}_{\text{axial}} = -2Cwxp \cos \theta + CLw \sqrt{\frac{2\rho}{g}} p \frac{dx}{dt} + CLwx \sqrt{\frac{\rho}{2gp}} \frac{dp}{dt} \quad (11.16)$$

The approximation is often made that the pressure drop across the valve is constant. If this assumption is made, $dp/dt = 0$, and the last term in (11.16) above vanishes. Equation (11.16) can then be written in the simple form

$$\mathcal{F}_{\text{axial}} = -K_1x - K_2 \frac{dx}{dt} \quad (11.17)$$

where K_1 and K_2 are constants. Thus, to a first approximation the Bernoulli force may be represented by an equivalent spring and coefficient of viscous damping as far as the valve-spool dynamics are concerned.

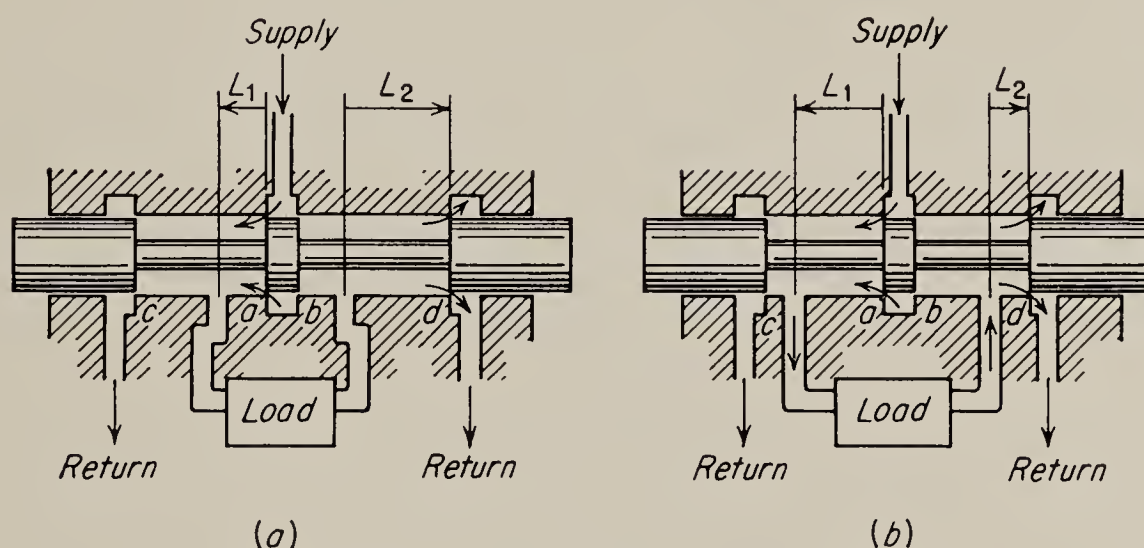


FIG. 11.16. Damping in hydraulic valves: (a) valve with positive damping; (b) valve with negative damping.

Since the steady-state force acts to close the valve the equivalent *Bernoulli spring* aids the centering springs usually employed with valves. The damping term may, however, be positive or negative depending on the sign of L . Usually a valve consists of several orifices in series, and the L 's for different orifices usually have opposite signs. Thus, consider the valve shown in Fig. 11.16a. In this valve L_2 is negative and L_1 is positive. Orifice a produces negative damping proportional to L_1 and orifice d produces positive damping proportional to L_2 . The net damping in the valve is therefore positive, since L_2 is greater than L_1 . In Fig. 11.16b is shown a valve for which L_2 is less than L_1 , and in this valve the damping from the hydraulic reaction force subtracts from the viscous damping. Since the hydraulic reaction force is a function of the pressure, it is therefore possible for the valve of Fig. 11.16b to exhibit negative damping for sufficiently high pressures. This may, of course, result in instability and violent oscillation of the valve.

Valves have been designed that counteract the effect of the steady-state axial force developed by the orifice.¹ Figure 11.17 shows the configuration of a force-compensated valve and a detail of one of the ports. In this device the inlet line to the chamber is constricted. The flow, entering at an angle θ_1 , tends to close the valve as usual. The chamber of this valve is shaped like a turbine bucket, and the flow leaving at an angle θ_2 is in such a direction as to generate a reaction force that tends to open the valve. In addition, the circulating flow entering the chamber at θ_3 also aids in keeping the valve open. While the design of this valve is essentially empirical, Lee and Blackburn report a reduction in the steady-state axial force by a factor of 50 compared with a conventional valve, and they state that the redesigned valve is essentially perfectly compensated throughout its operating range. The small notch shown dotted in Fig.

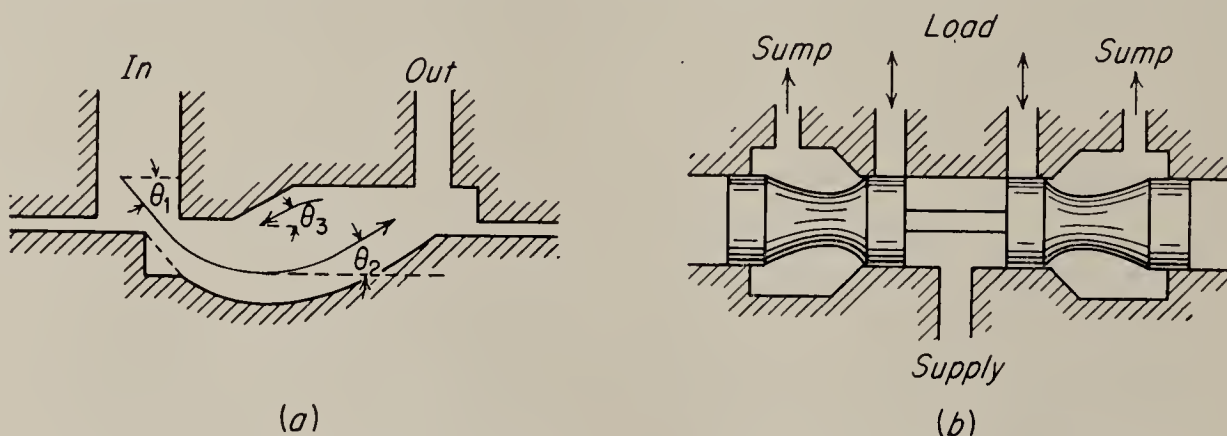


FIG. 11.17. Force-compensated valve. (From Lee and Blackburn)

11.17a simplifies the manufacture of the spool and does not appreciably affect the compensation.

In addition to special shaping of the valve lands, there are several other solutions to the problem of reducing axial forces on the valve spool. A two-stage valve may be designed to solve this problem, or an entirely different operating principle may be employed. Both of these possibilities are discussed below.

11.6. Radial Hydraulic Forces in Spool-type Valves. Radial or lateral forces on valve spools, if not symmetrical around the periphery of the spool, can cause the spool to move radially and lock or freeze against the cylinder wall. This problem of hydraulic lock has become more acute with the high pressures and low clearances of modern hydraulic systems.

Hydraulic lock may arise from several sources. One possible cause of hydraulic lock is the radial component of the Bernoulli force discussed in the previous section. In order to eliminate this possibility, the valve ports must be made symmetric around the valve periphery. A second possibility is dirt or metal chips that lodge in the radial clearance between piston and cylinder, thus wedging the piston. This possibility can be minimized by filtering the oil in the system.

¹ *Ibid.*

The most common source of hydraulic lock is a radial force produced hydrodynamically. Figure 11.18 shows a simplified view of a piston and cylinder. If the piston is a perfect right-circular cylinder and if the walls of the cylinder are parallel to the piston, there will be no net radial force due to the leakage flow. For any portion of the leakage flow, the clearance is constant, and the pressure gradient is uniform for the length of

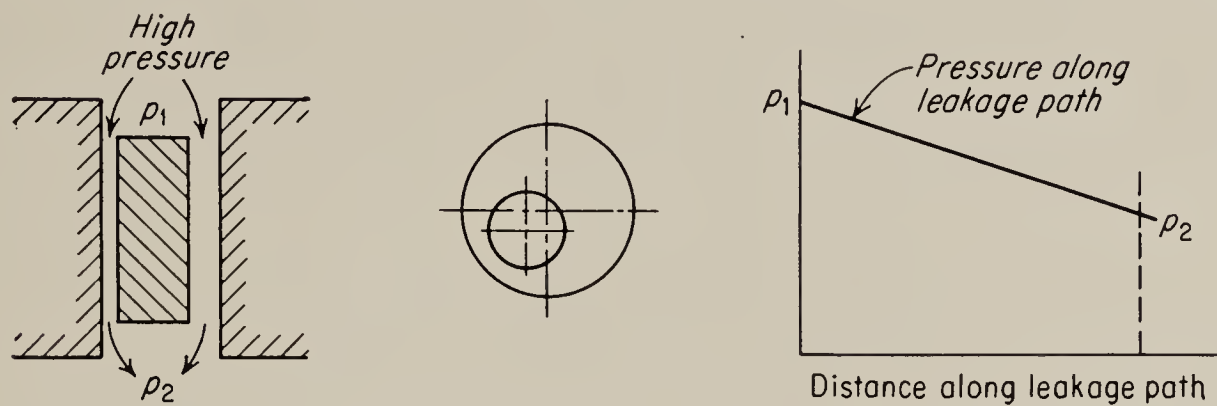


FIG. 11.18. Perfect right-circular cylindrical piston and cylinder, eccentric to one another.

the leakage path if we assume that the flow is laminar and that the velocity initially is zero. Thus even for an eccentric right-circular cylinder the radial pressures on the piston balance around the periphery of the valve.

The question naturally arises as to the forces involved if this perfect piston is somehow canted in the perfect cylinder, as shown in Fig. 11.19. If we consider two leakage paths l_1 and l_2 on opposite sides of the piston, the pressure distribution along the leakage paths will be as shown in Fig. 11.19*b* as a result of the interrelation of pressure head and velocity head. Thus throughout the length of the leakage path there will be a net pressure

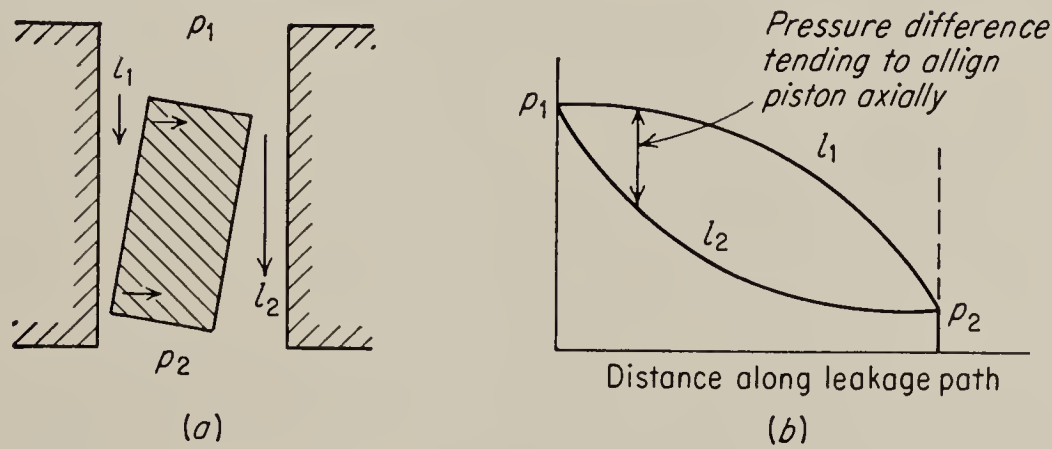


FIG. 11.19. Perfect right-circular cylindrical piston canted lengthwise in perfect cylinder.

difference tending to drive the piston to the right in Fig. 11.19*a* or, in general, toward the side with the constricted upstream flow. When the piston arrives at the cylinder wall, it will align itself with the wall, and the pressure drops along any leakage path will become equal. There will thus be no net force holding the piston against the cylinder, and hydraulic lock will not occur.

A somewhat more practical case than that of the perfect piston and cylinder is the case of a tapered piston in a perfect cylinder. Figure 11.20 shows an eccentric tapered piston. If the clearance is small, the leakage flow may be considered two-dimensional, and the relationship for laminar flow between two flat plates may be employed with little error to find the relationship between leakage flow and pressure.¹

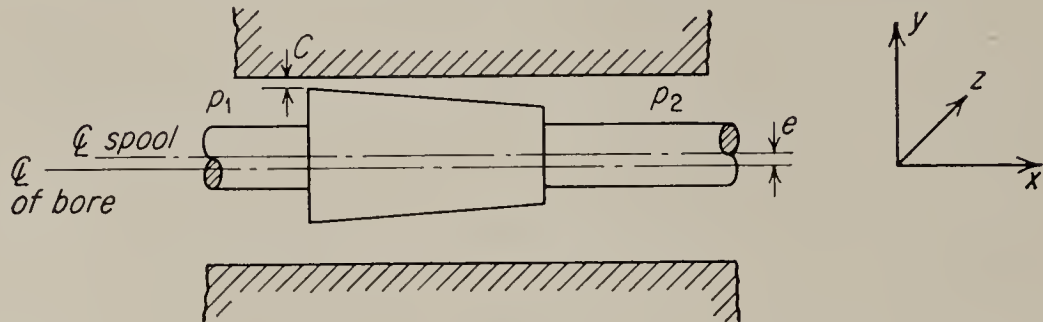


FIG. 11.20. An eccentric tapered piston.

By the same process² that is used to determine the Hagen-Poiseuille laminar-flow relation in a pipe, the average velocity of flow at any cross section of a leakage path may be determined as

$$V_x = - \frac{1}{3\mu} \frac{dp}{dx} \frac{y^2}{4} \quad (11.18)$$

where y is the distance between the plates. For an element of unit width the volumetric flow will thus be

$$\frac{F_x}{z} = - \frac{1}{3\mu} \frac{dp}{dx} \frac{y^3}{4} \quad (11.19)$$

where z is the width.

For a linear taper the relationship between y and x along a leakage path is

$$y = c + kx \quad (11.20)$$

where c is the clearance between piston and wall, as shown in Fig. 11.20, and k is the taper of the piston. Solving for p by integration,

$$p = \frac{6\mu F_x k}{z(c + kx)^2} + c_1 \quad (11.21)$$

The constant of integration, c_1 , can be evaluated by setting the pressure equal to p_1 where x equals zero. Then

$$p = \frac{6\mu F_x k}{z} \left[\frac{1}{(c + kx)^2} - \frac{1}{c^2} \right] + p_1 \quad (11.22)$$

Thus pressure is a parabolic function of x and varies with the taper k and clearance c . A smaller c results in a lower pressure at any station along

¹ D. C. Sweeney, Preliminary Investigation of Hydraulic Lock, *Engineering*, vol. 172, pp. 513-516, 580-582, 1951.

² Dodge and Thompson, *op. cit.*

the leakage path, as shown in Fig. 11.21. Thus there is a net radial pressure along the path in such a direction as to reduce the clearance. The result is hydraulic lock.

It may be seen from Eq. (11.22) that, if the taper k is negative, the net force will be in such a direction as to center the piston. For this reason, since some tolerance must necessarily be accepted in any manufacturing process, pistons are sometimes deliberately manufactured with a slight negative taper in order to eliminate the possibility of a positive-tapered piston being produced from a nominally zero-taper design. Quite often, however, this solution to the problem of hydraulic lock results in excessive leakage flow.

It has been found by Sweeny¹ and others that, if several radial grooves are cut into the surface of the piston land, to eliminate the pressure differ-

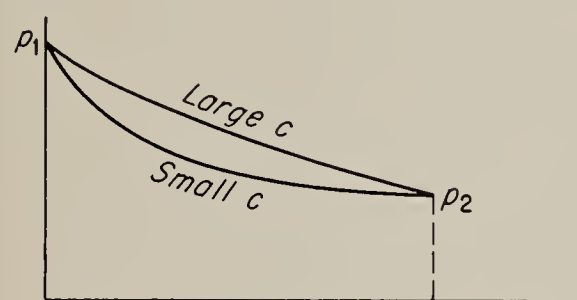


FIG. 11.21. Pressure drops in leakage paths of tapered piston.

ential around the spool, hydraulic lock can be eliminated. The size of the grooving should be large with respect to the clearances involved. Both the depth and width of the grooves should be at least ten times larger than the clearance if the clearance is of the order of a few ten-thousandths of an inch, in order to permit free flow around the periphery. Sweeny found

that five or six of these grooves reduce the locking force to about 1 per cent of its previous value in a typical spool.

Forces due to factors other than the pressure gradient caused by leakage flow are also sometimes important contributors to hydraulic lock. These factors include static friction, collection of dirt, and metal-to-metal contact. Sweeny has found that these forces including hydrodynamic lock itself build up rather slowly, taking 4 to 5 minutes to reach maximum in some cases. A common solution to the static sticking problem is to agitate the pilot valve at some high frequency either mechanically or by superimposing an a-c component on the valve-driving signal. This agitation is called *dither* and must be at a frequency high enough to assure that the controlled elements are unable to follow the rapid motion of the pilot valve.

11.7. Graphical Analysis of Control Valves. For large input signals the nonlinearity of valve characteristics cannot be ignored. Analytical methods of treating the problem are possible, but graphical techniques appear more convenient. The similarity between the characteristic curves has led to the extension of vacuum-tube analysis techniques to hydraulic valves. In this section we shall consider several of these techniques in detail. It should be noted that these nonlinear analyses do not

¹ Sweeny, *op. cit.*

permit the evaluation of equivalent gains or time constants for the elements; thus the approximations considered in the following section, which do permit this, are more often employed than these more exact methods.

It is possible to construct load lines on the valve characteristics that relate flow and pressure, such as those shown in Sec. 11.4. A load line is the pressure-flow relation for the load that is driven by the valve. Since the valve and the load are in series, the same flow must exist in each, and their two pressure drops must sum to the supply pressure. Since the load line and the valve characteristic represent relations that must be simultaneously satisfied, the point of operation must occur where the two curves

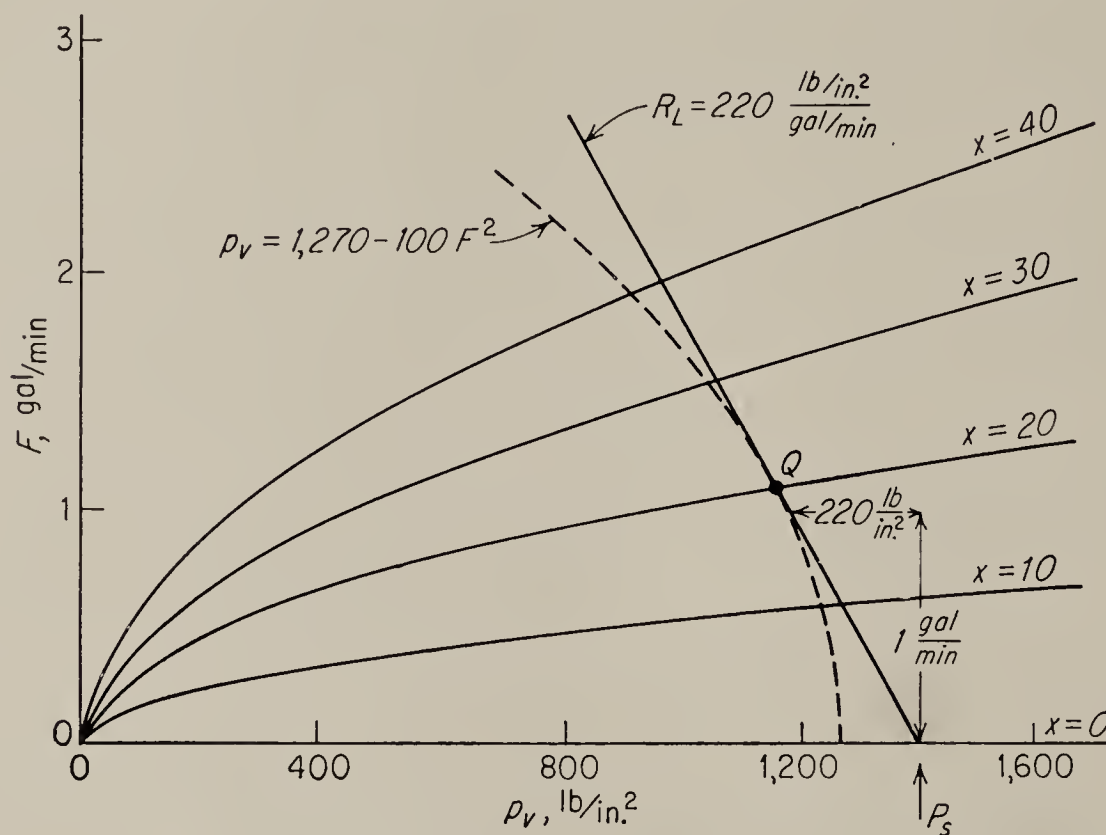


FIG. 11.22. A load line for a resistive load on an ideal valve characteristic. (Cunningham)

intersect. The load can be assumed to be so close to the valve that transmission-line effects can be ignored, or the effects of the line may be included in the load. Two load lines of this type are illustrated in Fig. 11.22. The valve characteristic curves on which these load lines are drawn are those of an ideal valve which has zero lap and square ports. The relation between pressure across the load (p_L) and flow through it is assumed linear for the solid line. The load line is constructed relative to the assumed supply pressure (p_s). The actual operating point of the system must lie on the load line at a point determined by the valve opening (x). If $x = 20$, the operating point Q will be as shown in the figure. Quite often the load presented to the valve has a nonlinear relation between pressure and flow. We might refer to this as a nonlinear hydraulic resistance. For instance, in Fig. 11.22 is shown dotted a parabolic load line such as would be presented by an orifice. If the

parabola passes through the operating point Q with the same slope as the linear resistance, it will have an equation

$$p_v = 1,270 - 100F^2 \quad (11.23)$$

It is possible to construct on the characteristic curves, lines of constant delivered power.¹ The power delivered to the load is proportional to the product of load pressure p_L and flow F .

$$W = p_L \left(\frac{\text{lb}}{\text{in.}^2} \right) F \left(\frac{\text{gal}}{\text{min}} \right) 231 \left(\frac{\text{in.}^3}{\text{gal}} \right) \frac{1}{12} \left(\frac{\text{ft}}{\text{in.}} \right) \frac{1}{33,000 \text{ ft/min}} \frac{\text{hp}}{\text{ft/min}} \quad (11.24)$$

or
$$W = 5.83 \times 10^{-4} p_L F \quad \text{hp} \quad (11.25)$$

where pressure is in pounds per square inch and flow is in gallons per minute. For a given supply pressure, curves of constant load horsepower

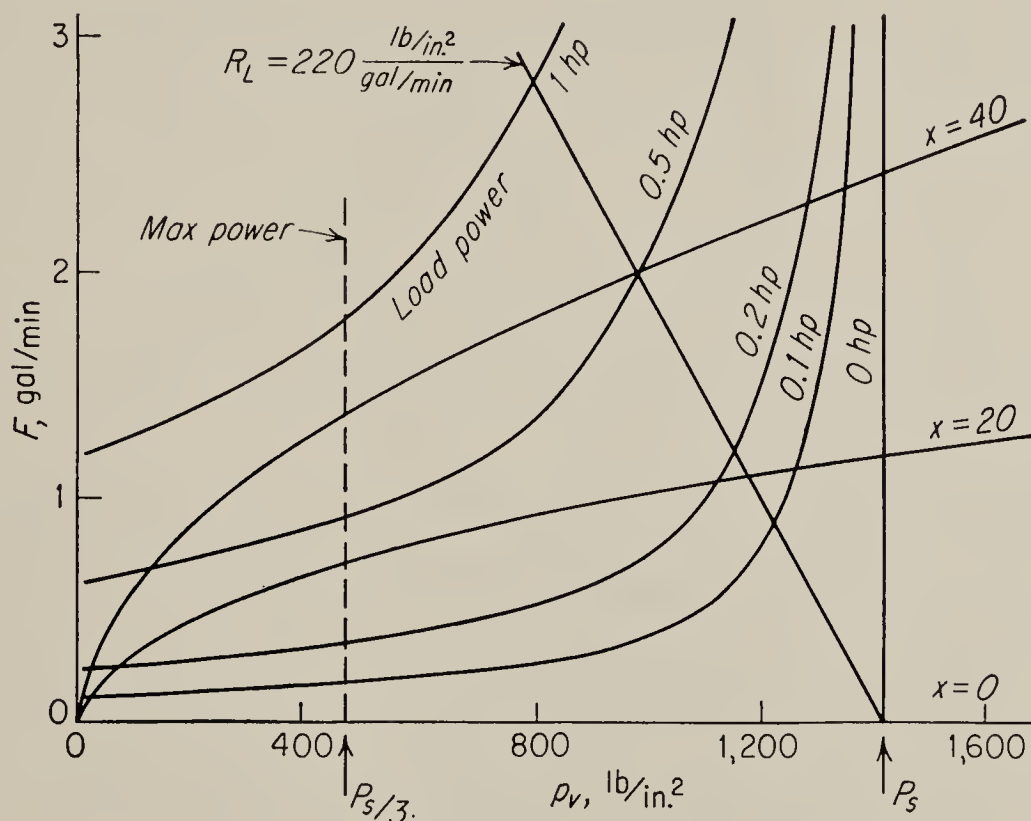


FIG. 11.23. Lines of constant delivered power superimposed on a valve characteristic. (Cunningham)

may be calculated from Eq. (11.25). These curves are hyperbolas and are shown plotted for p_s at 1,400 psi in Fig. 11.23. For a given value of x we can find the maximum power that can be delivered to the load and the pressure at which this occurs. From Eq. (11.12) we may retain the first term for a zero-lap valve and write for a constant x that

$$p_v = k_0 F^2 \quad (11.26)$$

and the load power may be written as

$$W = k_2(p_s - p_v)F = k_2(p_s - k_0 F^2)F \quad (11.27)$$

¹ W. J. Cunningham, "The Hydraulic Control Valve: An Analysis of Its Performance," Yale Univ. Dept. Elec. Eng. Tech. Rept., June, 1950.

Maximizing by setting dW/dF to zero yields $p_r = p_s/3$ and

$$p_L = \frac{2p_s}{3} \quad (11.28)$$

This contour or straight line of maximum power is shown in Fig. 11.23. In general the operating point should lie to the right of this boundary. All of these constructions that have been made above allow the designer to set the optimum region of operation for the valve and motor and to establish the response to a given input. A dynamic analysis may be made quite easily by moving the operating point along the load line in any manner prescribed by the input signal and picking off the required values of pressure and flow. A somewhat more elaborate procedure is required if the load is not a pure resistance.

In actual practice the pure-resistance load is quite rare. More often the load consists of resistance plus the inertia of the moving parts. Under these conditions the flow equation may be written

$$p_s = p_v + R_L F + J_L \frac{dF}{dt} \quad (11.29)$$

where J_L is the equivalent hydraulic inertia of the load. A methodical graphical technique may be employed to solve for the response of a system to various loads.¹

In this construction x is the actual piston displacement; thus reaction forces and time lags between the input to the torque motor or solenoid are not included. They must be taken care of in the analysis of the remainder of the control system. Let us assume that a step-function input signal is impressed on the valve and that we desire to determine the output as a function of time. If at some instant the flow has a value F_0 and an incremental change ΔF occurs, Eq. (11.29) becomes

$$p_s - R_L F_0 - p_v = \left(R_L + \frac{J_L}{\Delta t} \right) \Delta F$$

or

$$\Delta F = \frac{p_s - R_L F_0 - p_v}{R_L + J_L/\Delta t} \quad (11.30)$$

The numerator of this equation is the distance on the characteristic between the load line and the valve characteristic.

The construction necessary is shown in Fig. 11.24. It will be assumed that $p_s = 1,400$ psi and that, at $t = 0$, x is suddenly changed from 0 to 40. Assume that the hydraulic load has a resistance of 220 psi/gpm and an

¹ Cunningham, *op. cit.*, from which this example is adapted. Preisman, "Graphical Constructions for Vacuum Tube Circuits," McGraw-Hill Book Company, Inc., New York, 1943.

inertia of 3.11 (sec-lb/in.²)/gpm.¹ The load line of the hydraulic resistance is drawn on the static characteristic as before. Δt is chosen as 0.001 sec for convenience, and the operator $R_L + J_L/\Delta t$ is evaluated as 3,330 psi/gpm. At zero time there is no flow, even though the valve is open; thus $R_L F_0$ and p_v are zero. Thus from Eq. (11.11), the full supply pressure is effective in causing an increment in F . Next, from the point $p_v = p_s$ on the horizontal axis, a line is drawn with a slope of $R_L + J_L/\Delta t$. The intersection of this line with the static characteristic gives the flow F_1 at the end of the first interval of time. This point is located on the R_L

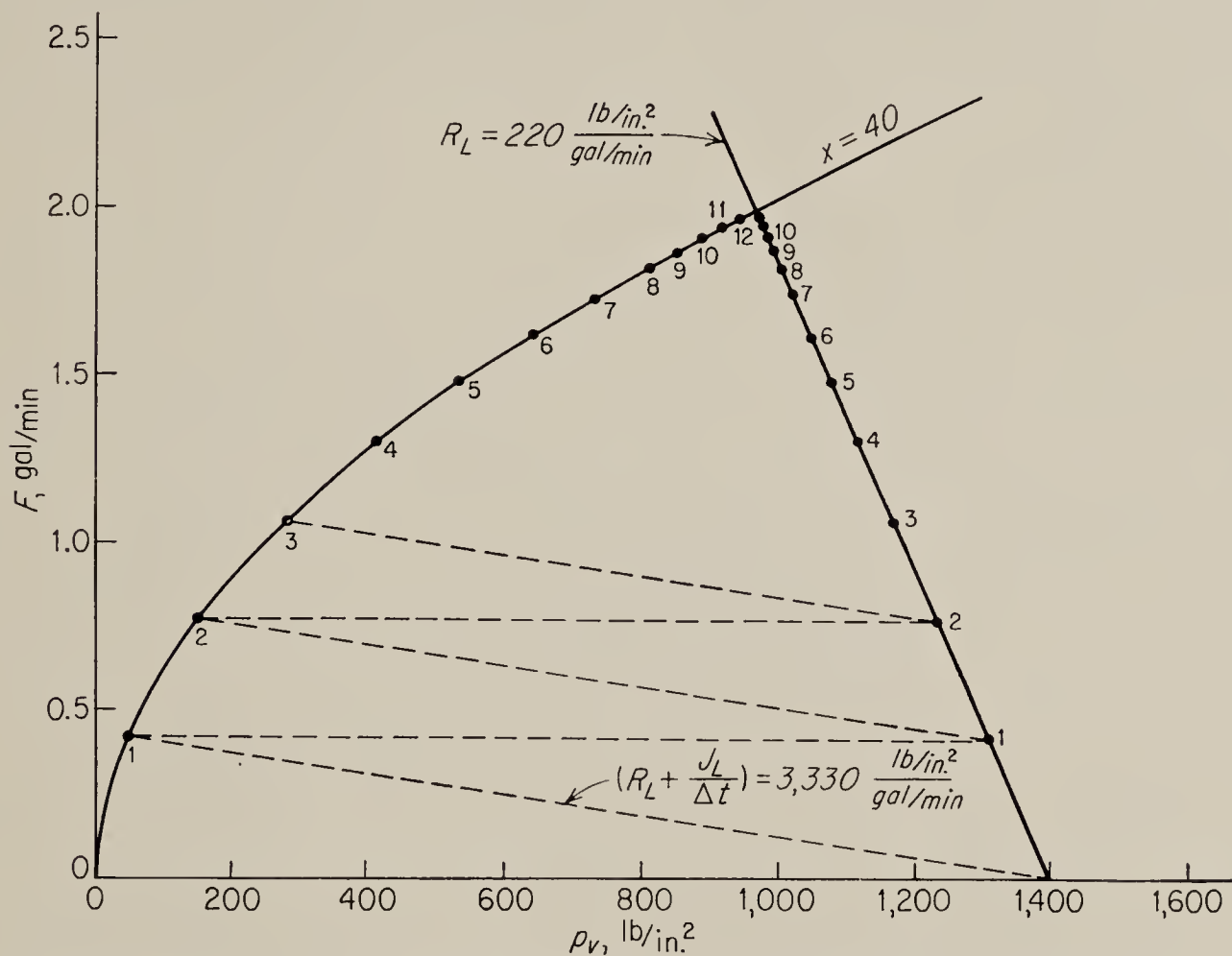


FIG. 11.24. Construction for valve supplying a load with inertia. (From Cunningham)

curve, and a new line at the same slope locates the flow at the second interval. The major approximation implied here is that the rate of change of flow is constant throughout the time interval chosen. Accuracy is improved by decreasing the time interval, but of course this requires more computation. With our choice of 0.001 sec for the time interval, about 12 construction steps are required to arrive at the operating point at the intersection of the load line and the characteristic curve. In Fig. 11.25 is shown this response plotted against time. The response for a linear approximation for the valve characteristic is shown dotted in the same figure for comparison.

¹ This corresponds to a hydraulic motor with a displacement of 0.38 in.³ per revolution with a mechanical load consisting of a resistance of 0.0174 ft-lb-sec and an inertia of 2.5×10^{-4} slug-in.³ per revolution.

The linear approximation was calculated from the equation

$$p_s = (R_v + R_L)F + J \frac{dF}{dt} \quad (11.31)$$

which has a solution

$$F = \frac{p_s}{R_v + R_L} \left[1 - \exp \left(-\frac{R_v + R_L}{J} t \right) \right] \quad (11.32)$$

and

$$p_L = p_s - FR_v \quad (11.33)$$

In these equations R_v represents an equivalent resistance for the valve. It seems reasonable to choose R_v so that the final value of flow will be the

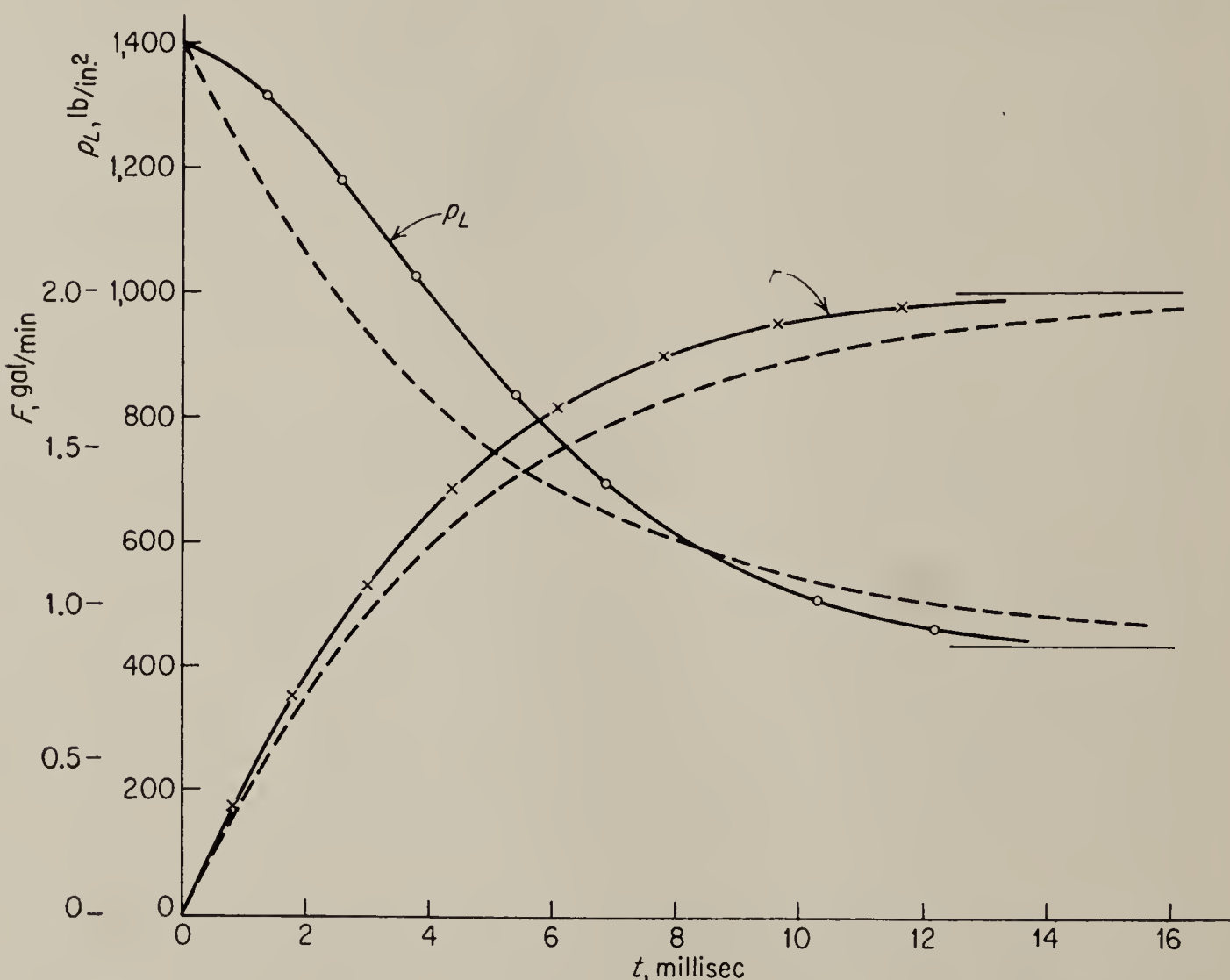


FIG. 11.25. Response of valve and load with inertia to a step input. (*From Cunningham*)

same for the linear approximation and the more accurate representation. The value of R_v is 482 psi/gpm under this assumption. We see from the figure that the flow increases more rapidly for the nonlinear construction, thus implying that the effective resistance of the valve orifice is less at small openings than would be assumed in the linear analysis.

It is possible to extend this construction technique to more involved inputs. For example, consider a sinusoidal driving signal in x . Let

$$x = \sin (2\pi 10t) \quad (11.34)$$

Since the values of x will be both positive and negative, we must employ a composite characteristic curve. This second family of static curves is plotted below the horizontal axis with the p_v scale reversed and the p_s points coinciding.

Let us suppose that it is desired to calculate the values of flow at 10° intervals in the sinusoidal variation; thus

$$\Delta t = \frac{10}{(360 \times 10)} = 0.00278 \text{ sec}$$

We must then construct characteristic curves for values of x at 10° intervals in ωt . The solution is begun as before by erecting the R_L line, as shown in Fig. 11.26. The value of the operator $R_L + J_L/\Delta t$ with the

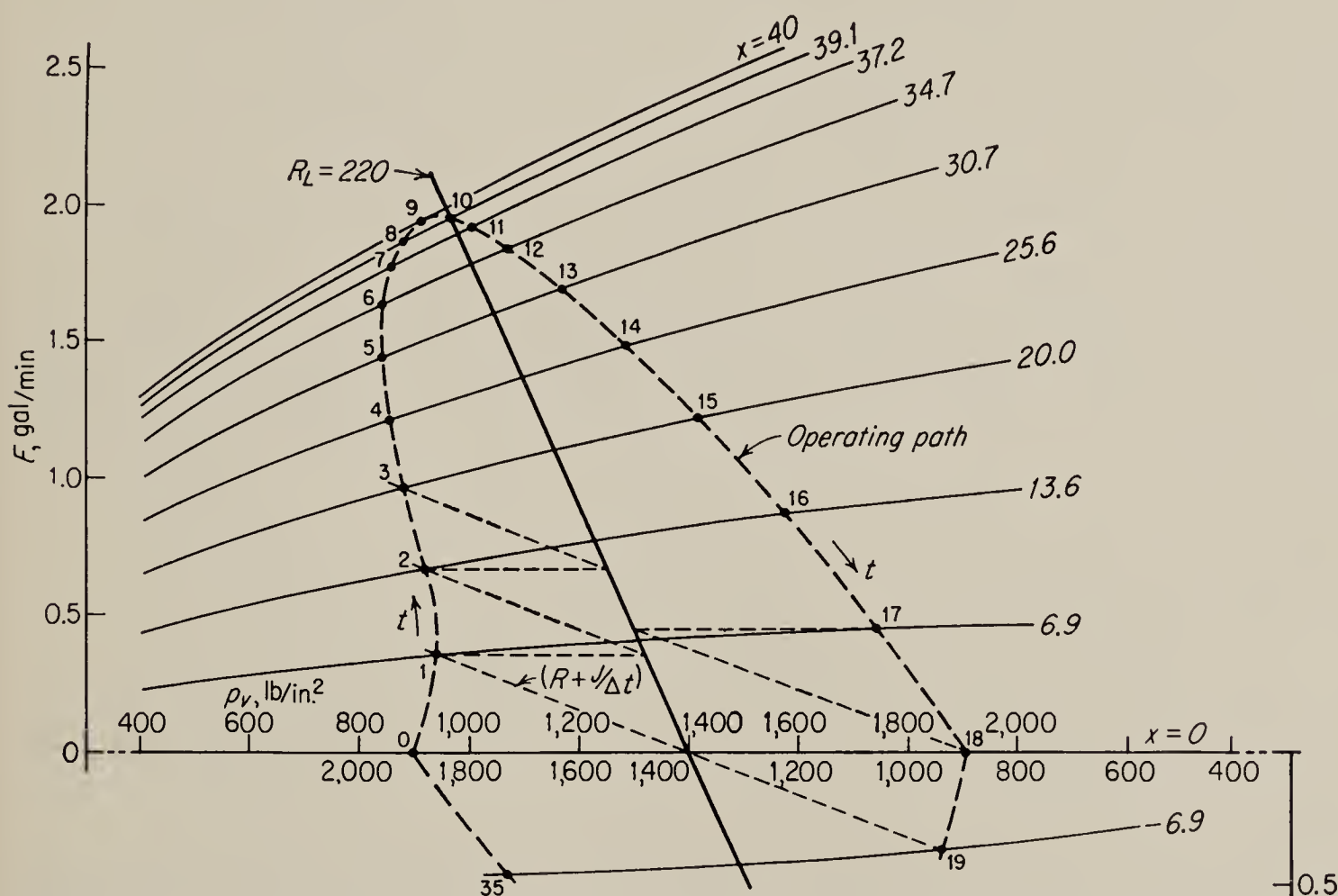


FIG. 11.26. Construction for inertial load and sine-wave driving function. (From Cunningham)

assumed value of Δt is 1,340 psi/gpm. A line is drawn from p_s at this slope, intersecting the static characteristic that applies at the end of the first time interval. The construction continues as before, employing, however, the particular static characteristic of interest at the particular moment of time. A curve may be drawn through the successive points, defining the operating path for the system. This type of solution is perfectly general and will include both the transient and the steady state. In this particular example, the path closes at each cycle, and the two parts of the solution cannot be distinguished. The solution may be

considered as a transient, repeated each half cycle. This is so, since in a zero-lap valve the flow must be zero whenever the stroke is zero. The plot of the various quantities of interest in this example is shown in Fig. 11.27 as functions of time. The flow is a distorted sine wave which lags the wave of pressure, as would be expected in a load with inertia. The small cusp in the pressure wave, which is due to the zero lap, would probably not be observed in practice because of practical effects such as leakage.

It has been pointed out that the nonlinear analysis does not yield a time constant or gain constant for the hydraulic valve. Indeed, by the

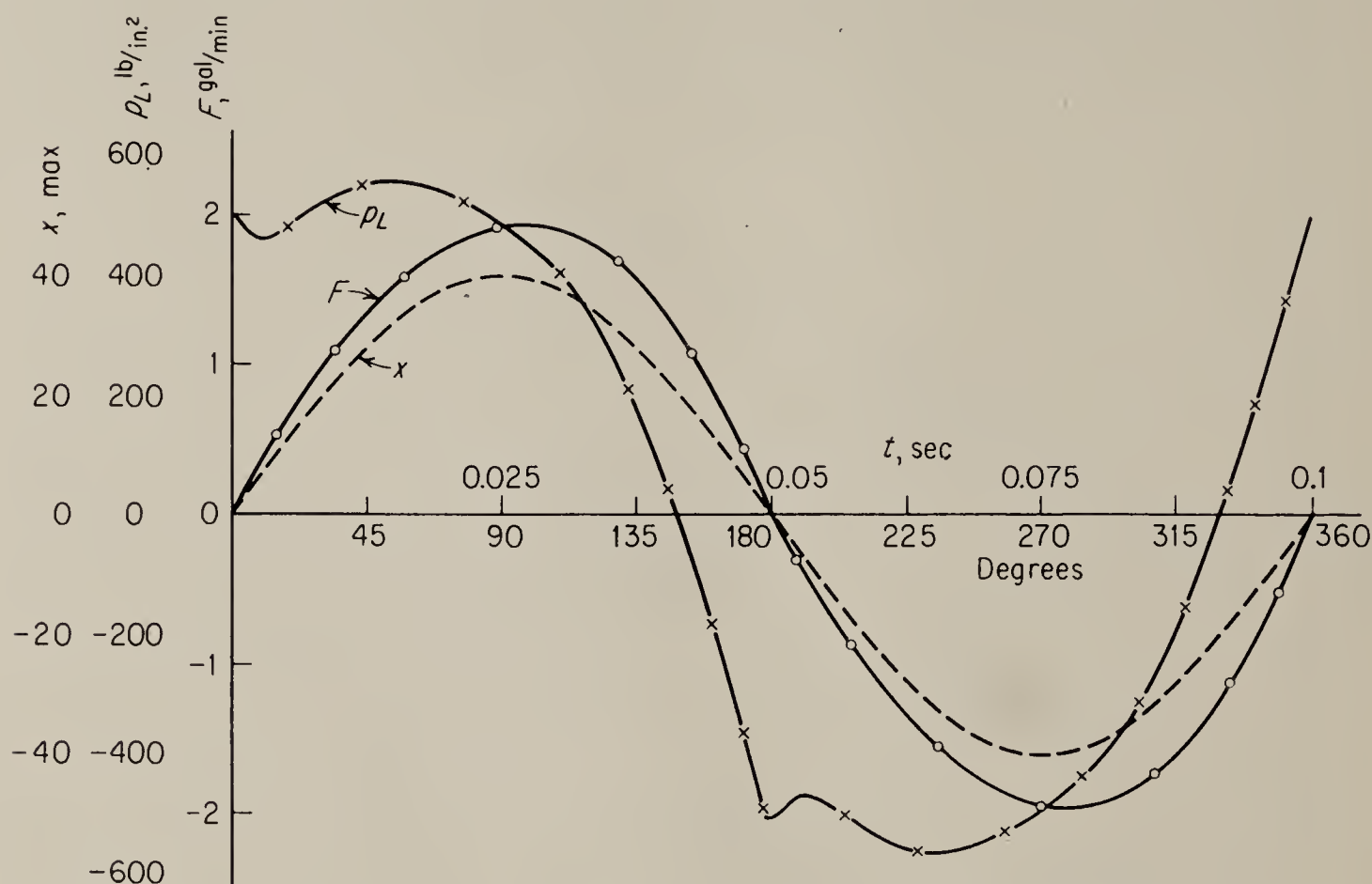


FIG. 11.27. Time response of a system with inertia for a sinusoidal driving function. (From Cunningham)

very definition of a nonlinear element, its transfer function depends on the input. However, it would be instructive to compare particular values obtained in the examples above, for any possible correlation. From the step-function analysis shown in Fig. 11.25 one can deduce a time constant of 4 msec and a gain of approximately 0.05 gpm per unit of stroke. Figures from the sinusoidal analysis do not compare well with these because of the rather unrealistic assumption of zero lap. The reader may show that at 20 cps the peak of the flow curve lags the x curve by about 15° , whereas a linear system with a time constant of 4 msec would have a phase lag of about 26° . If a more elaborate analysis were undertaken which included underlap, the correlation would be improved.

11.8. Linearized Small-signal Analysis. While the graphical analyses considered in the previous section yield more accurate results under given

conditions than a linear approximation, they are rather long and complicated, and they do not permit extension to general inputs. Under conditions of small driving signals it is possible to make good linear approximations with resultant simplification of the analysis.

The first linear approximation is suggested by an examination of Fig. 11.12, the flow-pressure relations for an underlapped valve in the underlapped region. The variation is almost linear in this region. The system is analogous to an electric bridge circuit, as shown in Fig. 11.28.

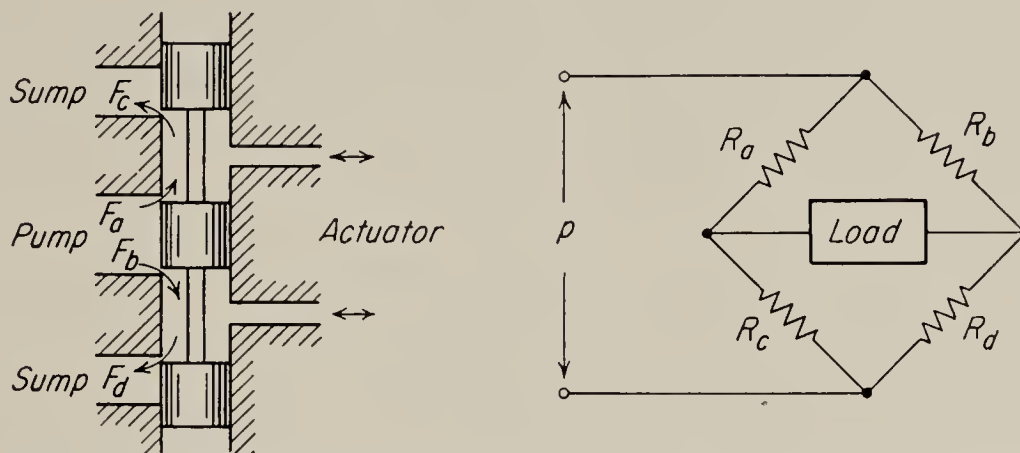


FIG. 11.28. Bridge analogue for underlapped valve in the underlapped region.

If the lapping is symmetric, the bridge is balanced when the valve is centered, and no fluid flows in the load. Motion of the spool down increases resistances R_a and R_d and decreases R_c and R_b . We can represent this by writing

$$\begin{aligned} R_c &= R - x \\ R_a &= R + x \\ R_b &= R - x \\ R_d &= R + x \end{aligned} \quad (11.35)$$

where R is the orifice resistance when the spool is centered and x is a parameter proportional to stroke. From Fig. 11.28 it may be seen that

$$R = \frac{p_s}{F_{\text{leak}}} \quad (11.36)$$

where p_s is the supply pressure and F_{leak} is the total leakage flow with the piston centered. The reader may show that

$$F_L = \frac{x p_s}{R R_L + R^2 - x^2} \quad (11.37)$$

Thus for small values of x we have the linear transfer function

$$\frac{\hat{F}_L}{\hat{x}} = \frac{p_s}{(R R_L + R^2)} \quad (11.38)$$

A more important application of linearization procedures may now be made for operation of a valve in general. The reader may recall that

vacuum tubes are also basically nonlinear devices, since the plate current as a function of plate or grid voltage is given by Child's law:

$$i_p = c(\alpha e_g + e_p)^{3/2} \quad (11.39)$$

However the analysis of vacuum-tube circuits proceeds with little difficulty. As is the case with vacuum tubes, hydraulic valves may be analyzed either graphically as above, when large-signal behavior is required, or on a small-signal "linearized" basis.

To simplify the linearization procedure, we assume that the valve is symmetrical, that there is no leakage in the load, that supply pressure p_s is constant, and that the sump pressure p_o is zero; i.e.,

$$F_1 = F_2 = F_L \quad (11.40)$$

$$p_1 - p_2 = p_L$$

$$p_1 + p_2 = p_s \quad (11.41)$$

$$p_o = 0 \quad (11.42)$$

where F_L is the load flow and p_L is the pressure across the load, and where other symbols are defined as in Fig. 11.11. Equation (11.41) shows that, if p_1 goes up a certain amount, p_2 goes down by the same amount.

The procedure is based on the assumption that, for small variations of the variables about an operating point, the characteristics are essentially linear. Thus, consider the nonlinear function between flow, pressure, and valve displacement:

$$F_L = F_L(p_L, x) \quad (11.43)$$

Differentiation gives

$$dF_L = \frac{\partial F_L}{\partial x} dx + \frac{\partial F_L}{\partial p_L} dp_L \quad (11.44)$$

If the differentials are replaced by small increments of the variables, this may be written as

$$\Delta F_L = G \Delta x - Y \Delta p_L \quad (11.45)$$

or, dropping the Δ 's,

$$F_L = Gx - Yp_L \quad (11.46)$$

where $G = \partial F_L / \partial x$ and $Y = -\partial F_L / \partial p_L$. G and Y will be approximately constant if the variation of x and p_L is not excessive. This procedure is exactly analogous to the process of differentiating Child's law [Eq. (11.39)] with the result

$$\frac{\partial i_p}{\partial e_p} = \frac{3}{2} C (\alpha e_g + e_p)^{1/2} \triangleq \frac{1}{r_p} \quad (11.47)$$

and

$$\frac{\partial i_p}{\partial e_g} = \frac{3}{2} C \alpha (\alpha e_g + e_p)^{1/2} \triangleq g_m \quad (11.48)$$

As an example of a typical G , take the relation for an underlapped valve given in Eq. (11.11) and find $\partial F_L/\partial x$.

$$G = \frac{\partial F_L}{\partial x} = k \frac{\partial(x_o + x)}{\partial x} \sqrt{\frac{p_s - p_L}{2}} - k \frac{\partial(x_o - x)}{\partial x} \sqrt{\frac{p_s + p_L}{2}} \quad (11.49)$$

$$\text{Thus} \quad G = \sqrt{2} k [\sqrt{p_s - p_L} + \sqrt{p_s + p_L}] \quad (11.50)$$

Similarly the reader may compute Y and for given parameters establish the transfer function for the valve. Equation (11.50) holds only in the underlap region, and when the relation for the valve outside the underlap region is found, it will be noted that the gain drops by a factor of 2. For large-signal operation the high-gain region about center may sometimes be ignored.

Once having accepted the restrictions implied in the small-signal analysis, the operation of the valve with reactive loads of all sorts may be considered. As a matter of fact, we may go further and consider the effect of the mass and friction in the valve stem itself. To do this, we assume that the input to the valve is a force source of finite stiffness. In order to get a complete expression of the various reactions taking place in the valve, the hydraulic reaction force must be considered. This force has been discussed in Sec. 11.5 and was shown to be a function of pressure, displacement, rate of change of pressure, and rate of change of displacement [see Eq. (11.16)]. Thus using the symbol \mathfrak{F}_h for reaction force, we can say

$$\mathfrak{F}_h = \mathfrak{F}(p_L, \dot{p}_L, x, \dot{x}) \quad (11.51)$$

where $\dot{p} = dp/dt$, and $\dot{x} = dx/dt$. For small variations of the variables Eq. (11.51) may then be written in the form

$$\begin{aligned} \mathfrak{F}_h &= \frac{\partial \mathfrak{F}}{\partial p_L} p_L + \frac{\partial \mathfrak{F}}{\partial \dot{p}_L} \dot{p}_L + \frac{\partial \mathfrak{F}}{\partial x} x + \frac{\partial \mathfrak{F}}{\partial \dot{x}} \dot{x} \\ &= \lambda_1 p_L + \lambda_2 \dot{p}_L + K_h x + B_h \dot{x} \end{aligned} \quad (11.52)$$

The parameters λ_1 , λ_2 , K_h , and B_h are considered constant for small variations of the variables, and may be obtained by differentiation of Eq. (11.16). If measured characteristics relating the reaction force to pressure and displacement are available, the parameters λ_1 and K_h may be obtained from them. This procedure has the advantage that the actual force characteristics rather than characteristics for an ideal valve are used, but the parameters λ_2 and B_h cannot very well be obtained this way. A compromise solution might therefore be used by which λ_1 and K_h are determined from experimental curves and λ_2 and B_h from the ideal relation.

To determine the equation of motion of the valve stem, it is assumed that the applied force is \mathfrak{F}_i and that this force is opposed by the hydraulic reaction \mathfrak{F}_h , the force of accelerating the mass of the valve stem, that is, $M d^2x/dt^2$, the force of friction between the stem and valve body, that is, $B dx/dt$, and the spring force Kx . Hence the equation of motion may be written in Laplace transform form

$$\hat{\mathfrak{F}}_i - \lambda_1 \hat{p}_L - \lambda_2 s \hat{p}_L - K_h \hat{x} - B_h s \hat{x} = (Ms^2 + Bs + K) \hat{x} \quad (11.53)$$

We also have the relation

$$\hat{F}_L = G \hat{x} - Y \hat{p}_L \quad (11.54)$$

Furthermore there is a relation between p_L and F_L through the load. The hydraulic impedance Z_L , which may be reactive, may be defined as

$$\hat{p}_L = Z_L \hat{F}_L \quad (11.55)$$

These three expressions may be solved directly to yield F_L or p_L as a function of \mathfrak{F}_i , but a greater insight into the operation of the valve is afforded by constructing an equivalent circuit for the valve to satisfy these equations. Such a circuit is given in Fig. 11.29. Equation (11.54)

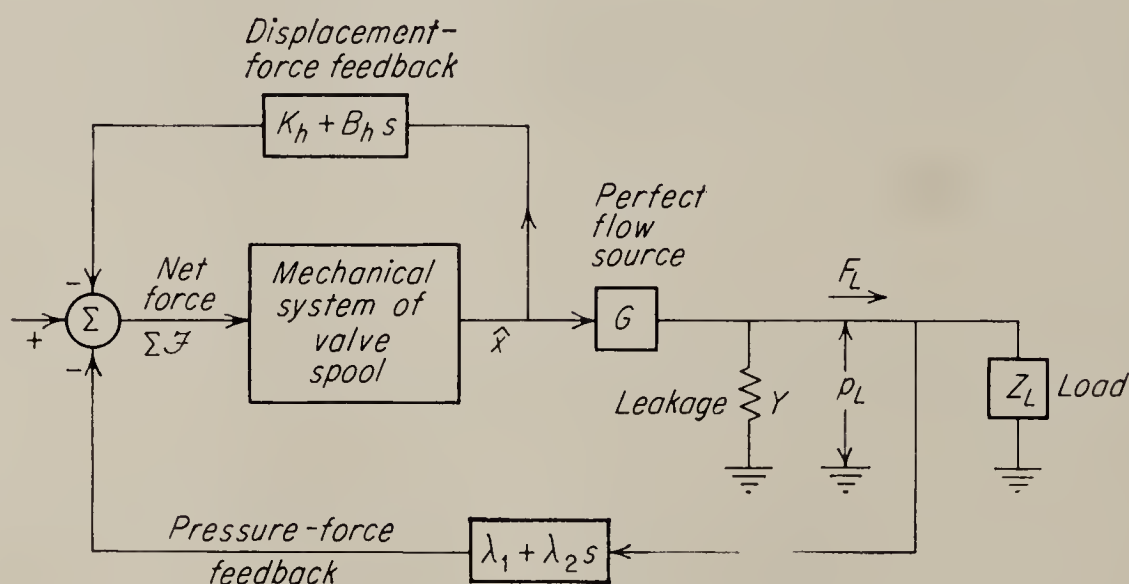


FIG. 11.29. Linearized equivalent of hydraulic-flow-control valve.

is represented by a perfect flow source yielding a flow F for a given input x . The fact that the actual valve is not a perfect flow source is taken into account by the leakage Y , which causes the flow to the load to be diminished when pressure exists. Reaction force is fed back from x and p_L as per Eq. (11.52) and results in a net force $\Sigma \mathfrak{F}_h$ tending to move the valve piston.

In most hydraulic actuators or motors the leakage and effect of compressibility are so small that the oil flow into them is a reliable index of output displacement. Hence, we may solve Eqs. (11.53), (11.54), and (11.55) for F_L , or we may obtain F_L as a function of \mathfrak{F}_i from the equivalent

circuit. The result in either case is

$$\hat{F}_L = \frac{\hat{\mathfrak{F}}_i G \frac{1/Y}{1/Y + Z_L}}{Ms^2 + B's + K'} \quad (11.56)$$

$$1 + \frac{G(\lambda_1 + \lambda_2 s)}{Ms^2 + B's + K'} \frac{Z_L/Y}{Z_L + 1/Y}$$

$$= \frac{\hat{\mathfrak{F}}_i G}{(Ms^2 + B's + K')(1 + YZ_L) + GZ_L(\lambda_1 + \lambda_2 s)} \quad (11.57)$$

where $B' = B + B_h$, and $K' = K + K_h$. It was shown in Sec. 11.5 that B_h is negative for valves for which the distance between the load outlets and the sump orifices is less than the distance between load outlets and pressure orifices (see Fig. 11.16). Therefore B' may be a negative quantity and the valve may be unstable and oscillate no matter what load is connected to it. However, if the load outlets are midway between the pressure and sump orifices, i.e., if $L_1 = L_2$ in Fig. 11.16, B_h and λ_2 will be zero. Under these conditions the valve may still be unstable if the load contains an inertia component so that $Z_L = R + Ls$. With this type of load the denominator of Eq. (11.57) will be a cubic polynomial of s . In order to determine whether or not the valve is stable the standard Routh test¹ may be applied to the denominator. For the case of $Z_L = R + Ls$, this test requires that for stability

$$B'M(1 + YR)^2 + B'K'Y^2L^2 + (B')^2YL(1 + YR) + B'YG\lambda_1L^2 + MGL\lambda_1 > 0 \quad (11.58)$$

The inequality shows that if $\lambda_1 = 0$ and $B' > 0$, the valve is stable. On the other hand, if $\lambda_1 < 0$, and $B' = 0$, the valve is always unstable. Since the parameter λ_1 is negative in a valve in which the hydraulic reaction force is not compensated it is apparent that a certain amount of positive damping ($B' > 0$) is required for stability.

We have seen here how a linear analysis can reveal valve instability. By examining Eq. (11.58) the designer may adjust the various parameters of the system to eliminate this possibility. The linear analysis has the advantage not only of relative simplicity but also of revealing the approximate transfer function of the unit. The accuracy of the analysis is reduced, however, if large input signals destroy the assumed linear relations about the quiescent operating point.

In this connection it should be noted that certain types of nonlinearities will cause oscillations in spool-type pilot valves under conditions which would allow the valve to be perfectly stable if the system were linear. To take one example, there is evidence to indicate that static friction is

¹ Chestnut and Mayer, "Servomechanisms and Regulating Systems Design," John Wiley & Sons, Inc., New York, 1951, Vol. I.

responsible for a mode of oscillation that is commonly observed in hydraulic and pneumatic valves. Another serious cause of instability is that in many cases the valve stem is not held securely by the valve-centering springs. Under these conditions the axial hydraulic reaction forces cause the stem to bounce from one spring to the other at a high frequency, giving rise to a chattering type of oscillation that will rapidly destroy motors and gearing connected to the valve.¹

11.9. Flapper Valves. One of the attempts to overcome the axial reaction force developed in piston valves has resulted in a modified design for the piston chamber, as discussed above. Further attempts to reduce this force and to reduce the high manufacturing costs of control valves have resulted in valves that operate on altogether different principles.

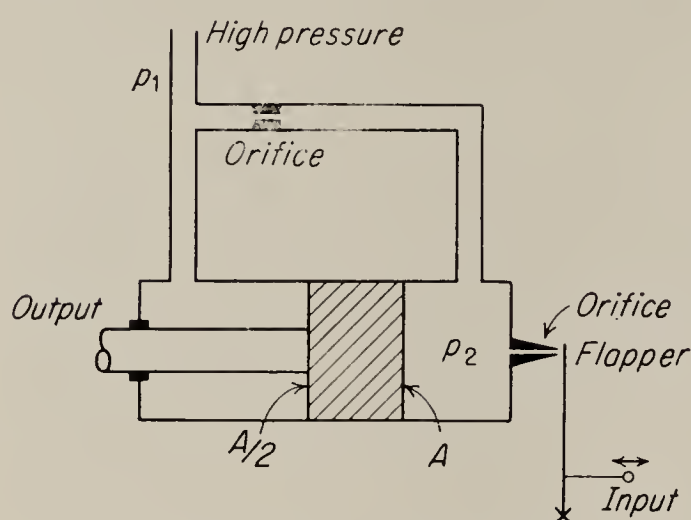


FIG. 11.30. Primitive flapper valve with unequal-area actuator.

Figure 11.30 shows one such device. The figure shows an elementary flapper valve. The high-pressure supply is connected to one side of the unequal-area valve piston. The pressure on the other side of the piston is controlled by varying the position of a flapper which partially closes a second orifice. The force that is developed on each side of the piston is the product of the pressure and the area across which the pressure is exerted. It will be

remembered that, although the pressure on the left-hand side of the piston is always higher than that on the right-hand side, the area of the left-hand side of the piston is smaller. Thus by properly adjusting the flapper, the forces may be balanced.

By application of the equations of flow through orifices the relationship between p_2 and flapper position may be found. From Eq. (11.7) the flow through the upstream orifice having a diameter D_1 is

$$F_1 = (\pi/4)CD_1^2 \sqrt{p_1 - p_2} \quad (11.59)$$

When the distance Y between the flapper and the orifice is very small the flow at the flapper is controlled approximately by the cylindrical area formed by extending the inner surface of the orifice pipe to the flapper. This area is $\pi D_2 Y$, where D_2 is the orifice diameter. Therefore, for very small Y we have approximately

$$F_2 = \pi C' D_2 Y \sqrt{p_2} \quad (11.60)$$

¹ J. L. Bower and F. B. Tuteur, Dynamic Operation of a Force Compensated Hydraulic Throttling Valve, *Trans. ASME*, vol. 75, p. 1395, 1953.

The discharge coefficient C' is probably different from the coefficient C used in Eq. (11.59) since the shapes of the orifices are different.

The flapper ceases to control the orifice when the cylindrical area $\pi D_2 Y$ referred to above approaches in magnitude the orifice area $(\pi/4)D_2^2$, or when $Y = 0.25D_2$. For very much larger Y the flow is controlled solely by the orifice and

$$F_2 = (\pi/4)CD_2^2 \sqrt{p_2} \quad (11.61)$$

From Eqs. (11.59), (11.60), and (11.61) it is possible to obtain a relation for the chamber pressure p_2 as a function of flapper displacement Y . For this purpose we ignore the difference in discharge coefficients and let $C' = C$. This is permissible if only an approximate relation is desired.

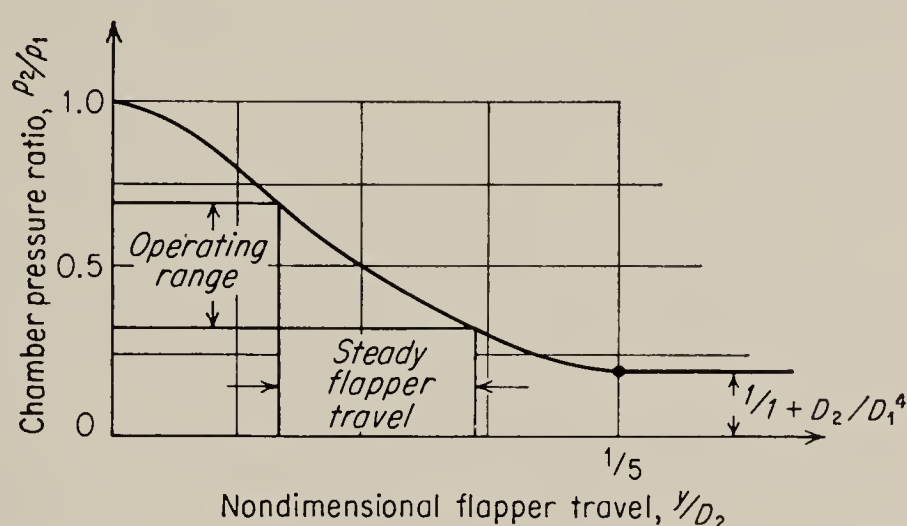


FIG. 11.31. Chamber pressure versus flapper travel, normalized. (From Nightingale)

Also, suppose that the piston is blocked so that the flow to the piston is zero. Under these conditions $F_2 = F_1$ and therefore for $Y/D_2 < 0.25$ we have

$$(\pi/4)CD_1^2 \sqrt{p_1 - p_2} = \pi CD_2 Y \sqrt{p_2}$$

or

$$\frac{p_2}{p_1} = \frac{(D_1/D_2)^4}{16(Y/D_2)^2 + (D_1/D_2)^4} \quad (11.62)$$

For $Y > 0.25D_2$ we equate Eqs. (11.59) and (11.61) giving

$$\frac{p_2}{p_1} = \frac{(D_1/D_2)^4}{1 + (D_1/D_2)^4} \quad (11.63)$$

A plot of chamber pressure normalized with respect to supply pressure versus flapper position normalized to orifice diameter is shown in Fig. 11.31.¹ Despite the nonlinear equation relating these variables it is seen that in the operating range the pressure is approximately a linear function of flapper position.

By equating forces on the piston and by performing the usual partial

¹ J. M. Nightingale, Hydraulic Servo-valve Design, *Machine Design*, vol. 27, no. 1, p. 191, 1955.

differentiations as in Sec. 11.8, it is possible to obtain a small-signal linear approximation for the flapper-control system. By equating flows in and out of the control chamber and writing a force balance for the piston, a transfer function \hat{x}/\hat{y} may then be obtained. The details of this computation, however, are left to the reader (see Probs. 11.10 and 11.11).

The hydraulic reaction force on the flapper may be approximated, if the stream is assumed to strike the flapper at right angles, by

$$\text{Force} = \frac{F\rho v}{g} \quad (11.64)$$

where F is the volumetric flow, v is the velocity of discharge, ρ is the density of the oil, and g the acceleration of gravity. The flow and discharge velocity can be found by the relations given above. This force can be considerably reduced by arranging two nozzles, opposing one another on either side of the flapper, as in the Moog valve discussed below.

The major limitation on the flapper valve is that the controlled-pressure chamber must be kept rather small if the time constant of the valve, which consists of the resistance of the upstream orifice and the capacity of the controlled-pressure chamber, is to be kept small. Enlarging the flapper orifice will improve the response of the valve, but at the expense of increased flapper reaction force and leakage flow. These restrictions place a limitation on the load that can be controlled by a flapper valve, and as a result the flapper valve is most often employed as the first stage of a two-stage valve.

11.10. Nozzle Valves. The Askania nozzle valve is a valve that has a very low hydraulic reaction force.¹ A schematic of the valve is shown in Fig. 11.32. The hydraulic flow is formed into a jet by the movable nozzle, and the flow is directed into one of two ports on either side of the receiving block and conducted to the load. The only force that must be supplied by the input is that required to overcome friction at the pivot and the inertia of the jet pipe. The major disadvantage of this device is its relatively high leakage. The frequency response of this device with the standard 8-in.-long, $1\frac{9}{32}$ -in.-diameter jet pipe peaks at about 30 cps and has a damping ratio (ζ) of 0.2.

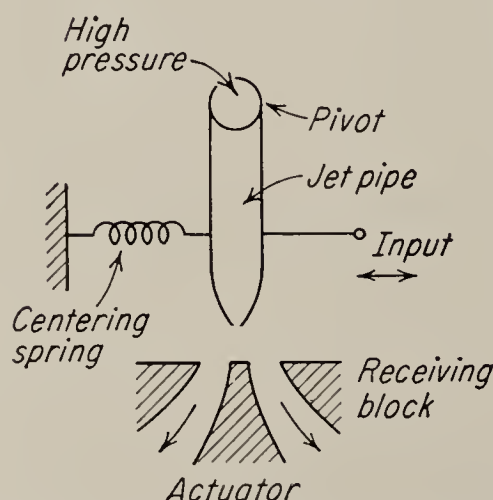


FIG. 11.32. Askania jet-pipe valve.

11.11. Slide Valves. Owing to the high cost of manufacturing spool-type valves to the very close tolerances required by high-performance systems and owing also to the high leakage of the flapper-type valves,

¹ S. Z. Dushkes and S. L. Cahn, Analysis of Some Hydraulic Components Used in Regulators and Servomechanisms, *Trans. ASME*, vol. 74, p. 595, 1952.

efforts have been made to develop a valve with neither of these disadvantages. The slide valve¹ shown in Fig. 11.33 is one possible answer. The advantage of the slide valve is that the slide and block can be clamped together, so that the two critical holes may be drilled through the block and slide at the same time. The spacing between the holes is not critical. Following this operation, the plugs and bushings are inserted to complete the critical portion of the device. Exact alignment is thus assured for all the orifices. The vertical spacing between slide and block may be established by a milling operation. The slide is suspended by bars and actuated by a torque motor of some sort, or the slide may be allowed to rotate with respect to the block. In either case, the variation of area of the orifices is essentially linear with respect to motion of the slide. The slide valve can be force-compensated in the same manner as spool-type

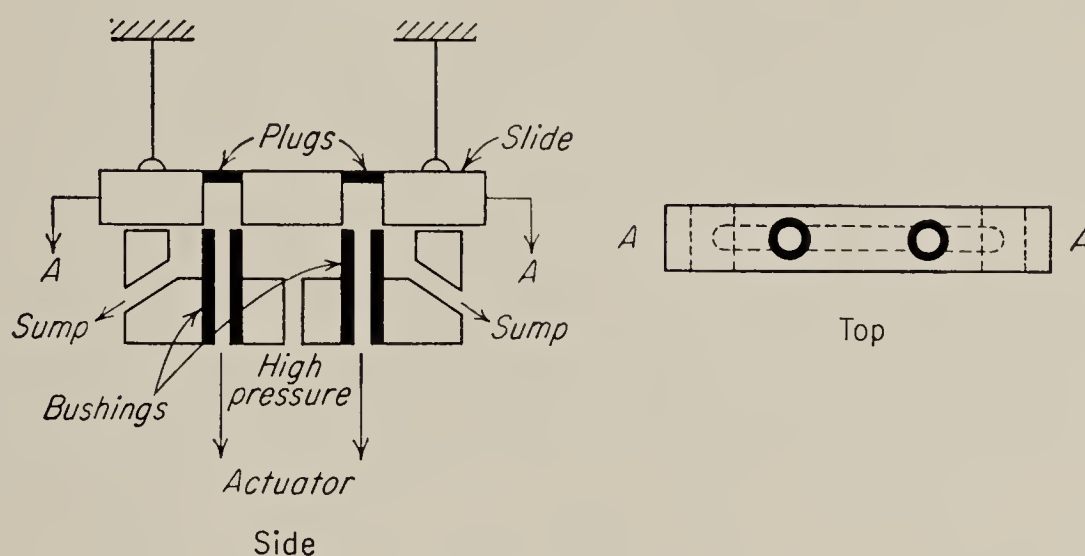


FIG. 11.33. A linear slide valve. (From Lee)

valves, or the valve with rotary motion for the slide plate may be designed so that the Bernoulli force of one nozzle cancels that of another.² In fact, in the configuration shown in Fig. 11.33, a component of hydraulic force acts to open the slide if the slide is displaced from center. Unfortunately, however, this opening force increases as the valve slide is opened, while the Bernoulli force decreases. Thus these two components cannot be relied upon to cancel each other. Figure 11.34 gives a plot of flow versus displacement of the slider of a rotary-plate valve. Note the linear characteristic, and note that the leakage flow is only about 1.5 per cent of full output flow, whereas the usual flapper-valve leakage flow is about 5 to 10 per cent of full-load flow. A practical disadvantage of the slide valve is the scoring of the bushings and block by chips and dirt that have a tendency to accumulate in the slide cavities.

¹S. Y. Lee, New Valve Configurations for High-performance Hydraulic and Pneumatic Systems, *Trans. ASME*, vol. 76, p. 905, 1954.

²*Ibid.*

An actuator that has been constructed¹ embodying the slide-valve principle is shown in Fig. 11.35. The ram could be used to position an output element mechanically, or it could be the spool of a high-power pilot valve designed for the same service as the two-stage piston valves discussed below and the flapper valves described above.

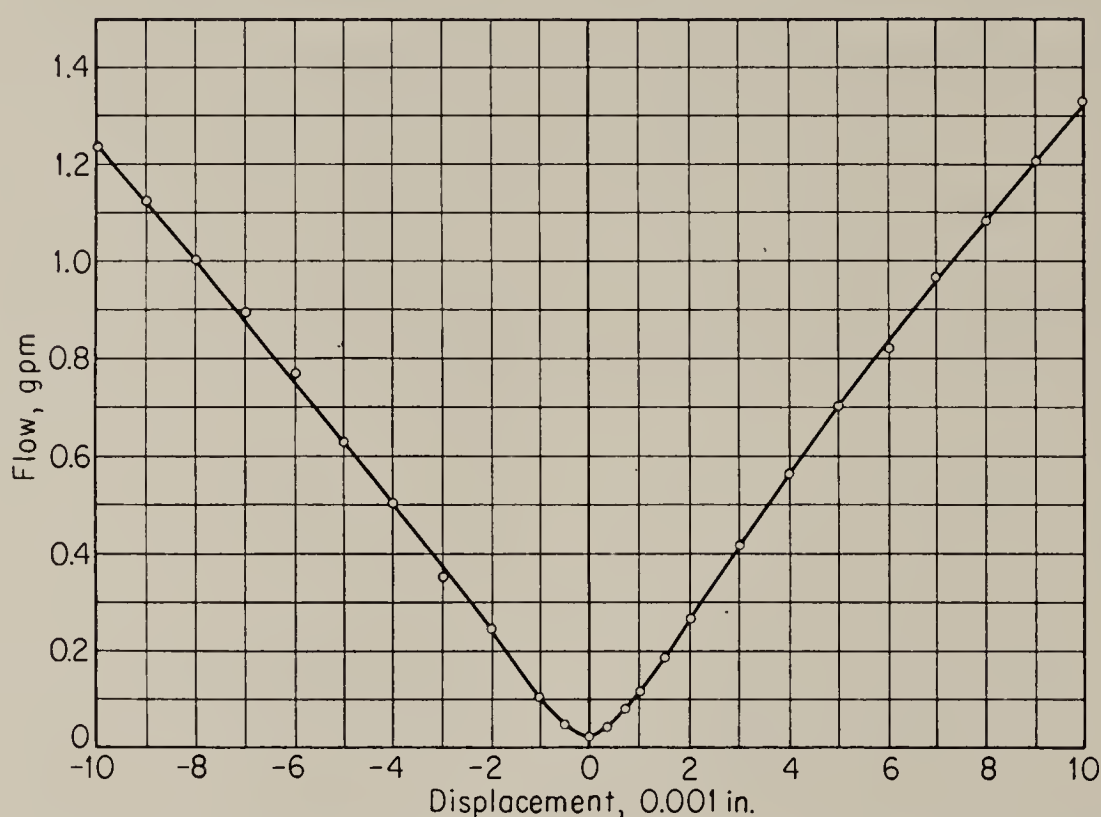


FIG. 11.34. Flow versus displacement for a rotary slide valve. (From Lee)

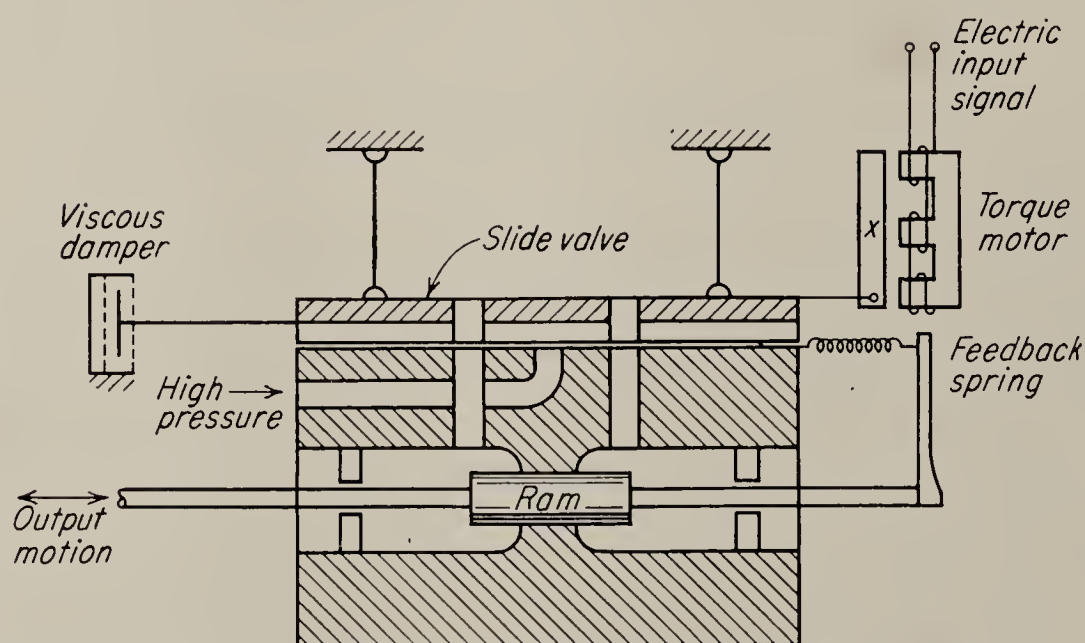


FIG. 11.35. Actuator employing a slide valve. (From Lee)

The frequency response of this actuator is given in Fig. 11.36. It will be seen that, in addition to its other advantages, the slide-valve principle allows a rigid, compact design with very high frequency response. Rotary slide valves of this type are presently available commercially.

¹ S. Y. Lee and J. L. Shearer, A Miniature Electrohydraulic Actuator, *Trans. ASME*, vol. 77, p. 1077, 1955.

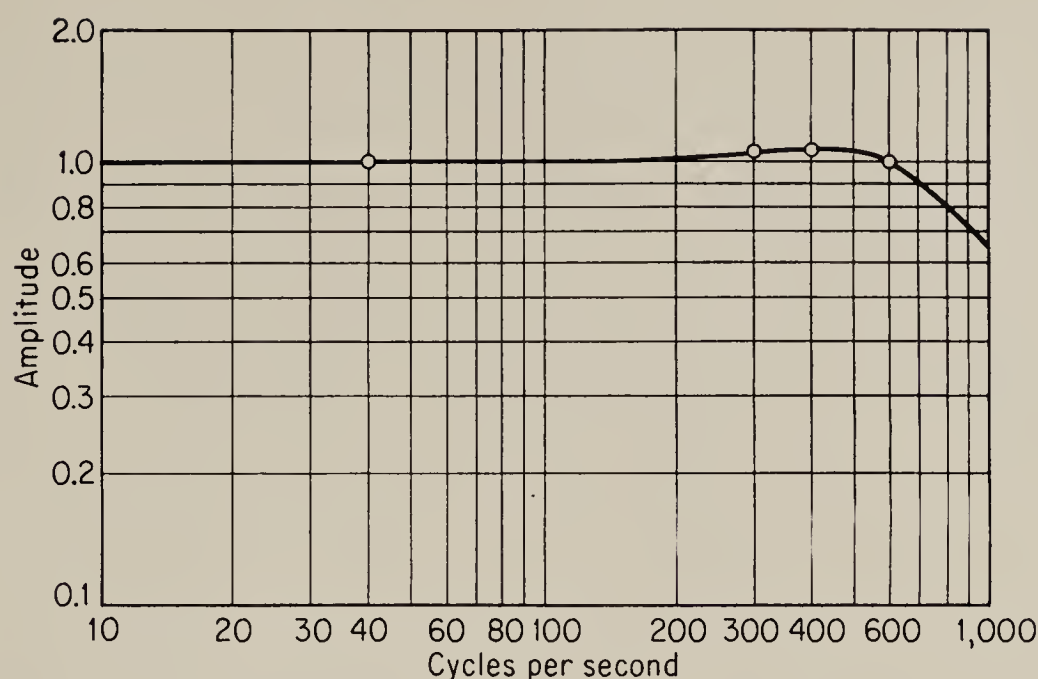


FIG. 11.36. Frequency response of actuator and slide valve. (*Lee and Shearer*)

11.12. Two-stage Valves. There are rather definite practical limitations on the power-handling capacity and the speed of response at these high powers for single-stage valves. Various combinations of the valves already discussed have been used in two-stage valves to overcome these limitations.

Let us first consider a two-stage spool-type valve, shown in Fig. 11.37. This valve has an operating characteristic somewhat different from that of a single-stage valve. A given displacement of the input results in a given flow into the actuator cylinder. Thus the actuator takes up some constant velocity. It may be seen, then, that the main valve opens at a constant rate and that the flow to the load constantly increases. The transfer function from input position to main fluid flow thus contains an integration. While this integration is sometimes desirable, it is usually removed by placing a feedback loop around the pilot valve and actuator cylinder.

Figure 11.38 shows three ways of adding feedback in a two-stage spool valve. In Fig. 11.38a the feedback is through a linkage. When the input moves down, for example, the pilot-valve spool also moves down, with the actuator cylinder as a fulcrum, thus porting high-pressure fluid to the actuator cylinder. As the cylinder moves down, the feedback link moves about the input as a fulcrum and moves the pilot-valve piston up, thus cutting off the flow of oil. In effect, then, the actuator follows the input.

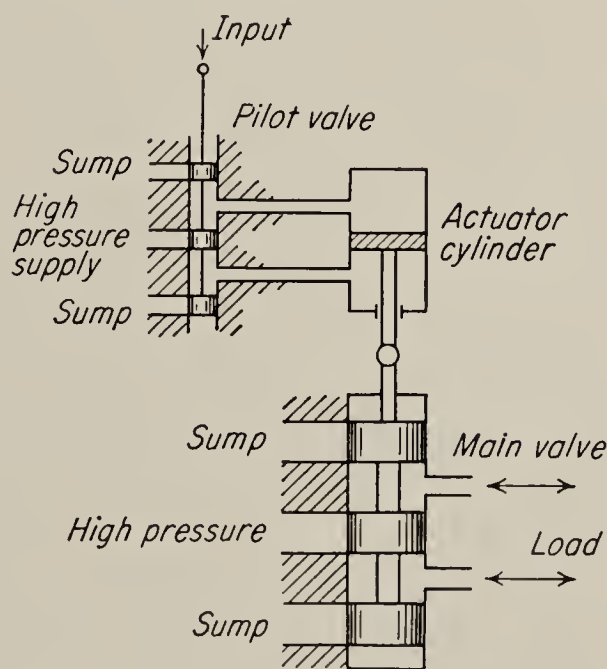


FIG. 11.37. Two-stage hydraulic valve.

In Fig. 11.38*b* the input varies the position of a movable sleeve in the pilot valve. If the input is moved down, for instance, the high-pressure oil is ported to the bottom of the actuator cylinder. The actuator moves up, carrying the pilot-valve piston down by means of the feedback link, thus stopping the flow of oil.

Figure 11.38*c* shows a Vickers two-land sleeve valve that operates in a manner somewhat similar to the action of Fig. 11.38*b*; however, the principle of differential area is employed in the system. In Fig. 11.38*c* the areas are in the ratio of 2:1, which is the most common arrangement. If the input is moved down, the valve spool ports high pressure to the bottom of the actuator cylinder, and the top of the actuator cylinder is opened to sump. The actuator will move up, and the sleeve will be

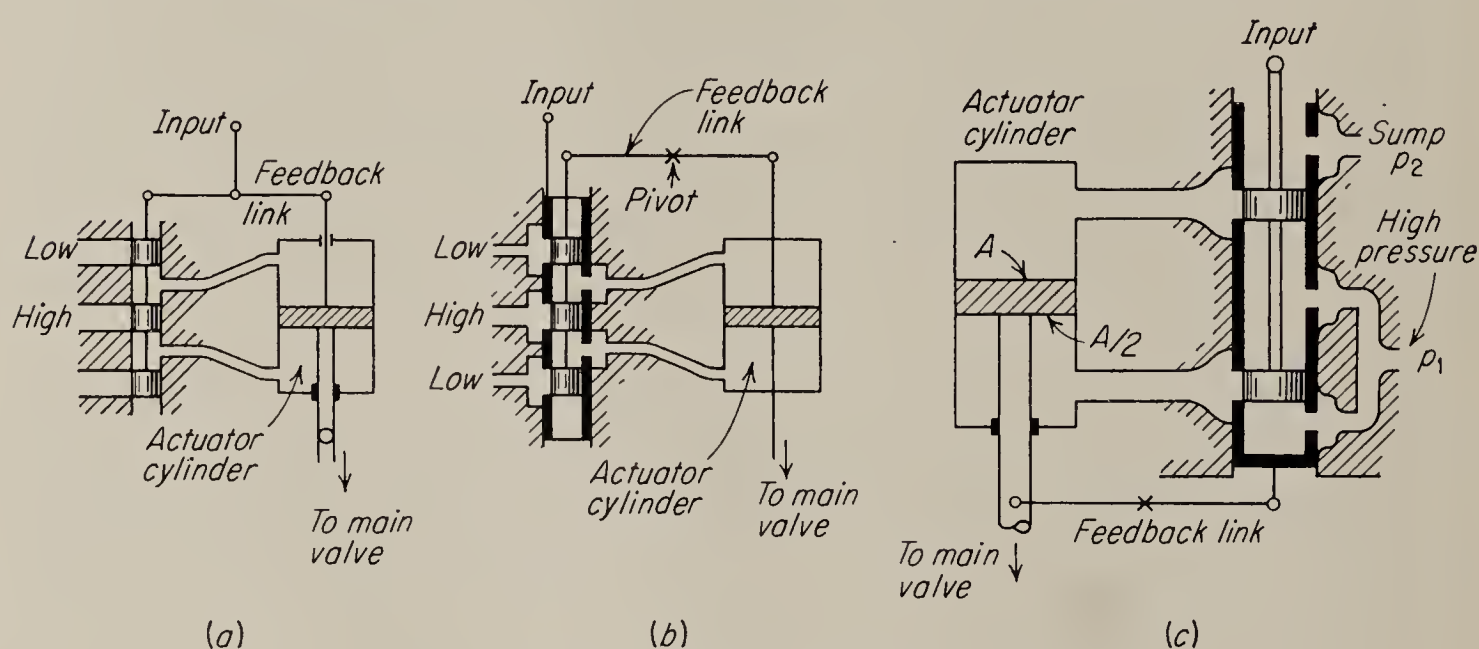


FIG. 11.38. Feedback links for pilot valve.

pulled down by the feedback link, thus closing off the fluid flow. The actuator follows the input with a reversal of direction. Now, if the input is moved up, a rather different effect occurs. The high-pressure fluid is ported to both the top and bottom of the actuator cylinder. The net force is down.

It is also possible to construct very compact two-stage valves without constructing the parts separately and connecting them by linkages. Figure 11.39 shows a two-stage slide valve used in a hydraulic autopilot control manufactured by Siemens (Germany) during World War II. If the force developed in the left-hand chamber, p_4A_4 , is equal to the force developed in the right-hand chamber, p_3A_3 , the spool is at rest. If the input is moved to the right, for instance, the orifice R_4 is reduced in size and p_4 increases, thus increasing the force in that chamber and moving the spool to the right. The new position of the spool will be such that the chamber forces are once more balanced. The motion of the spool ports oil to the main actuator. While this valve has been described¹

¹ R. Hadekel, Hydraulic and Pneumatic Servos, *Automation*, March, 1955.

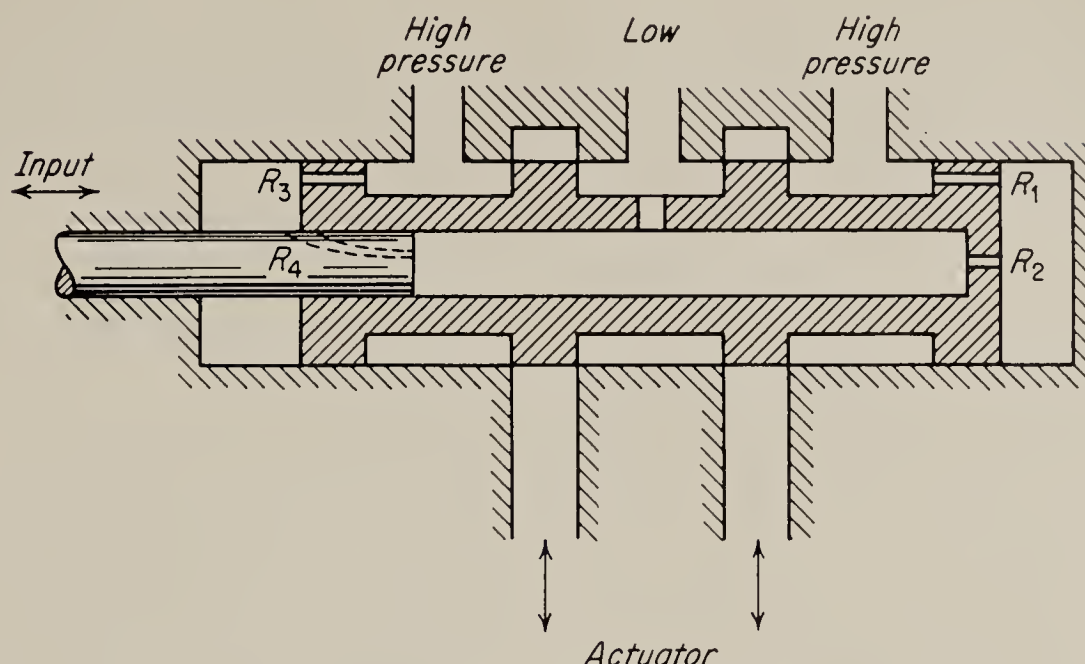


FIG. 11.39. A Siemens two-stage hydraulic pilot valve.

as analogous to a Wheatstone bridge with the resistance arms R_1 , R_2 , R_3 , and R_4 , it would appear to be more direct to consider it on the differential-area principle, as above, since in equilibrium p_3 is not equal to p_4 . The Pegasus valve of this type in which the input is supplied by a solenoid is presently available commercially.

A more compact two-stage valve is one that employs a flapper valve as its first stage. An elementary form of such a valve is shown in Fig. 11.40. The pressure p_2 , controlled by flapper position, actuates the spool against the spool spring. A disadvantage of this valve is that, for a given flapper position, a change in the high pressure will cause a change in the position of the spool, since the force developed by p_2 is opposed only by the spring.

The Moog¹ valve, shown in Fig. 11.41, is one of the earlier of the low-reaction-force two-stage hydraulic valves and is designed to operate at 1,000 to 3,000 psi. The arrangement of balanced nozzles at the flapper reduces the force required by the flapper, and the use of two chambers and two springs eliminates the direct reliance on the value of the supply pressure

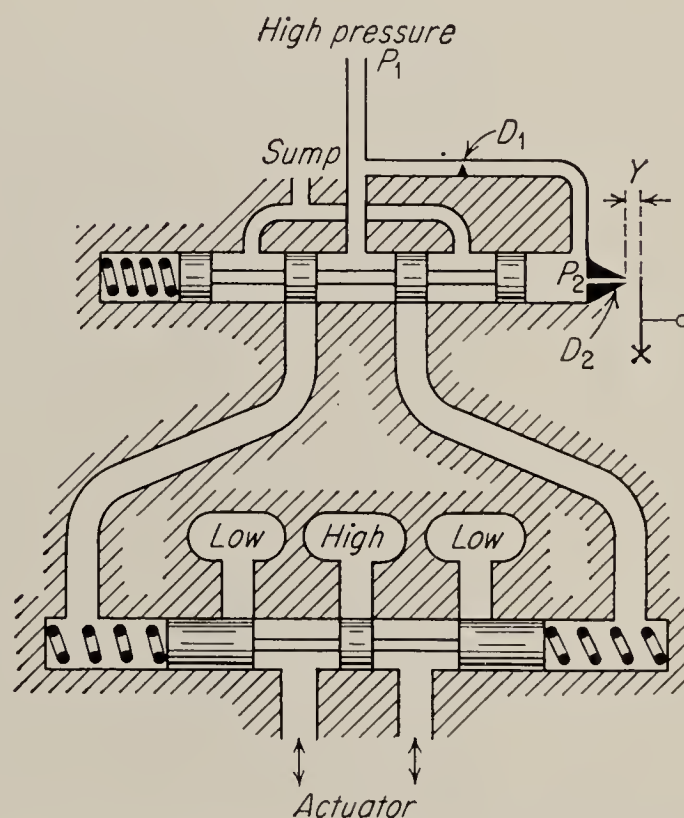


FIG. 11.40. Primitive form of a two-stage flapper-spool valve. More practical two-stage valves of this type have been developed.

¹ R. L. Scrafford, "Hydraulic Servos Incorporating a High-speed Hydraulic-amplifier Actuated Valve," *Trans. AIEE*, part II, vol. 72, p. 175, 1953.

characteristic of the primitive type. The force on the flapper is developed by the differential current flow in the two halves of the split-coil winding. Eight milliamperes differential current is sufficient to cause full rated flow of about 8 gpm in the valve. Valves capable of

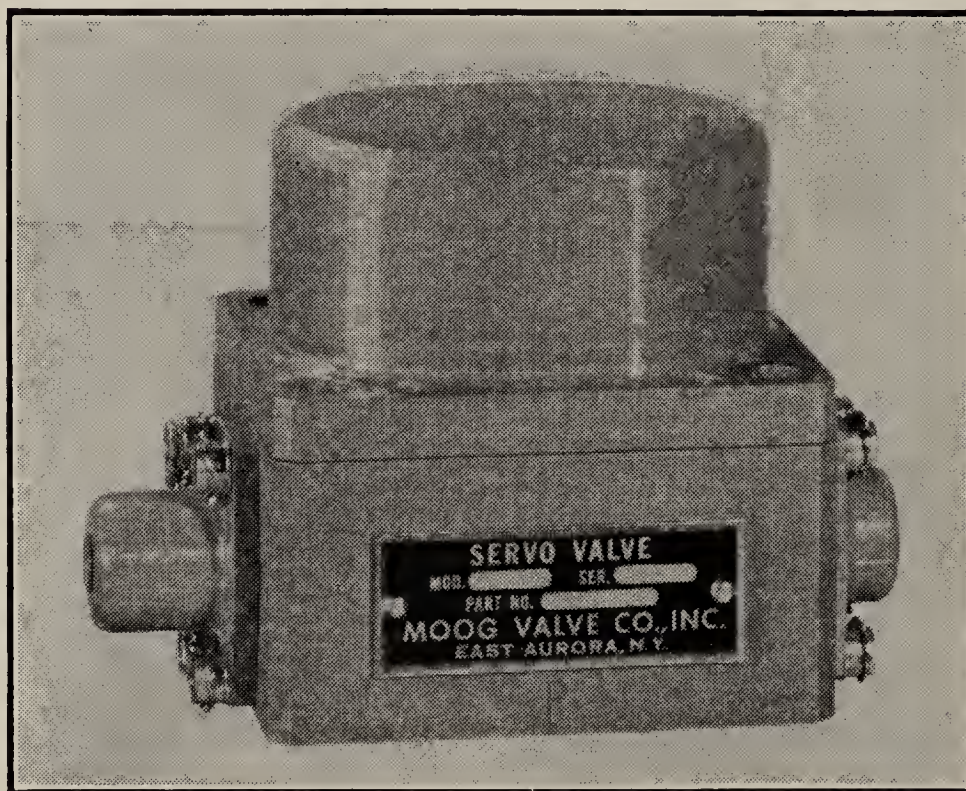
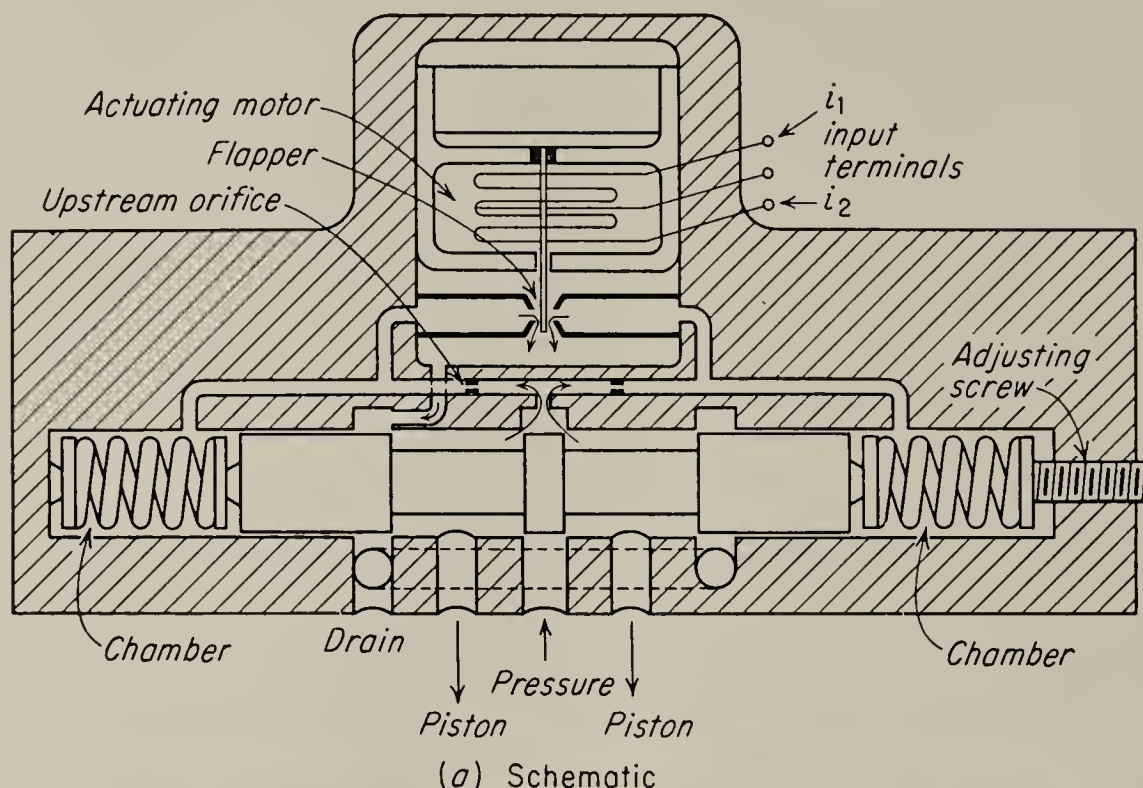


FIG. 11.41. The Moog valve.

larger outputs are also available. The leakage flow through the nozzles at zero signal is typically about 2 per cent of full rated flow for the 8-gpm model. This leakage flow cannot be reduced proportionally for smaller-flow models. From test characteristics furnished by Moog, it would appear that ζ , the damping ratio, is about 0.7, and the critical or natural frequency of resonance is about 40 cps for the Model 500. Thus

empirically

$$\frac{F}{i_1 - i_2} = \frac{1}{(M/K)s^2 + (B/K)s + 1} = \frac{1}{(s^2/6.3 \times 10^4) + (s/28.3) + 1}$$

(11.65)

The amplitude characteristic versus frequency is shown in Fig. 11.42. Test data for a model designed for a maximum flow of 0.5 gpm show a $\zeta = 0.5$ and a natural frequency of about 130 cps. The spring-mass

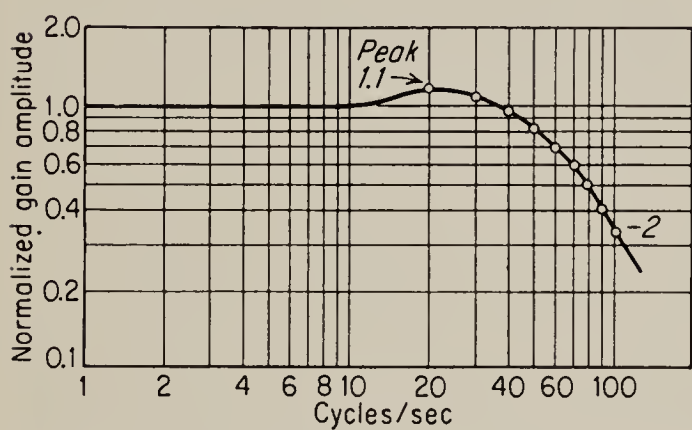


FIG. 11.42. α diagram for Moog Model 500 valve. (Courtesy Moog Valve Co.)

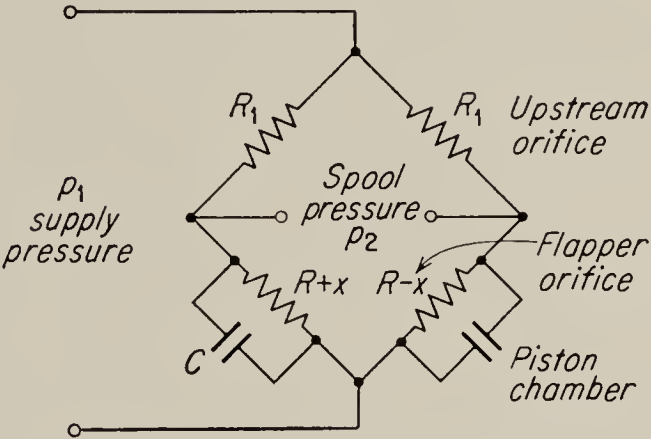


FIG. 11.43. Equivalent circuit of a portion of the Moog valve, where x is flapper position.

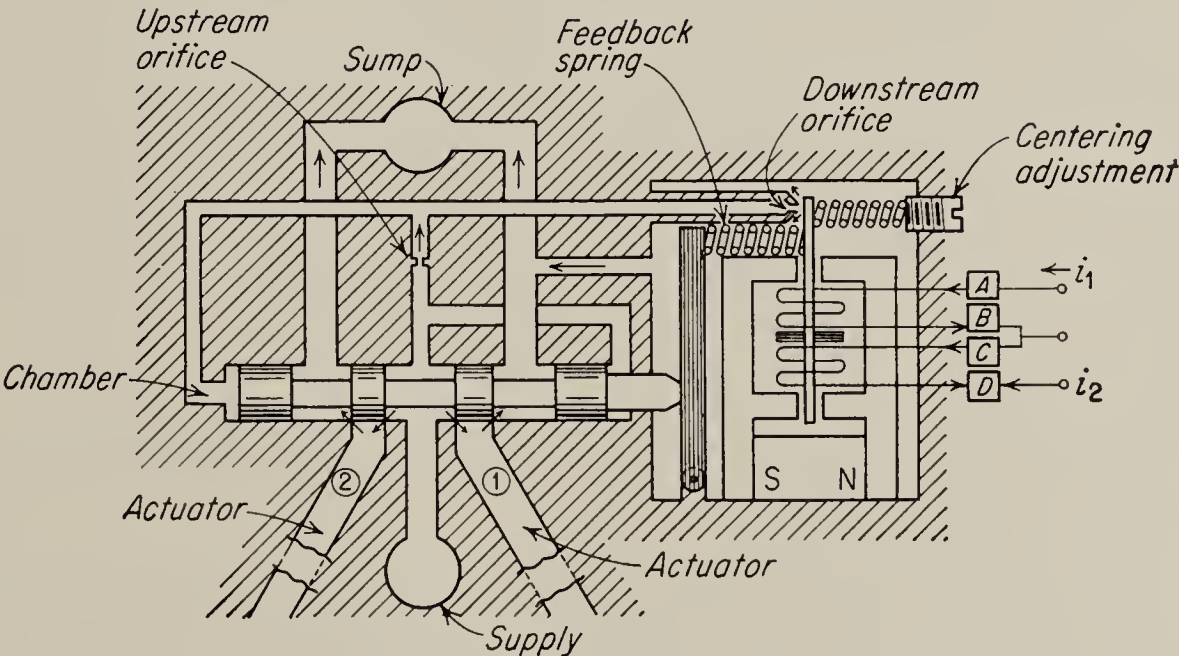


FIG. 11.44. Cadillac Model FC2 valve.

resonant frequency of the second-stage spool is about 1,000 cps for the typical Moog valve; thus it is well above the frequencies of interest. The valve may be thought of as the equivalent of a bridge circuit, as shown in Fig. 11.43. This circuit could not of itself cause the underdamped response shown in Fig. 11.42. The flapper is flexible, and the interaction between the nozzle flow and flapper position is perhaps responsible.

The Cadillac Model FC2 flow-control valve, as shown in Fig. 11.44, is a two-stage valve similar to the Moog valve. Unlike the Moog valve, however, the piston of the Cadillac valve is not directly restrained by

centering springs. The spool is operated on the differential-area principle by flapper-controlled pressure, and the piston position is fed back to the flapper through a feedback spring. This accomplishes the centering of the piston as effectively as spring centering. The time constant of the two actuating springs and the mass of the flapper are small enough to be

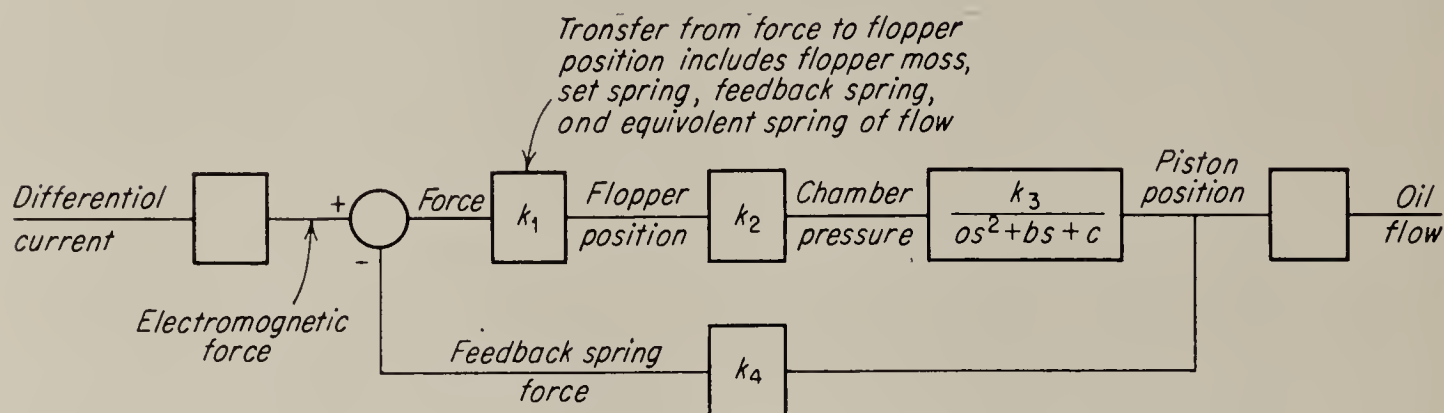


FIG. 11.45. Equivalent block diagram of the Cadillac Model FC2 valve.

negligible in this device. The time constant of the Cadillac FC2 valve is determined by the gain of the feedback loop. Figure 11.45 gives a block-diagram representation of the FC2 valve. The block diagram shown is only one of several possible block diagrams that could be drawn. The choice of force feedback is arbitrary. The flapper spring-mass system

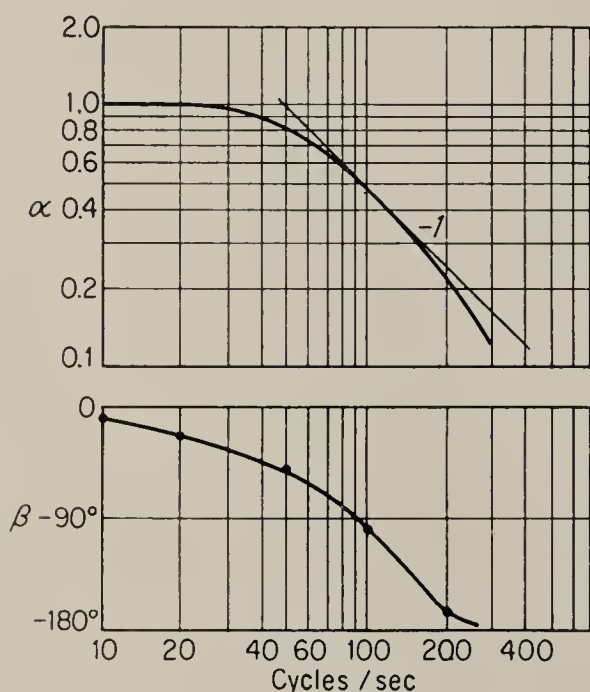


FIG. 11.46. Frequency response of Cadillac FC2 valve. (Courtesy Cadillac Valve Co.)

has a natural or critical frequency that is quite high and may be neglected. Likewise the transfer function from piston position to feedback spring force may be considered a constant. Thus the transfer function of the closed loop from input to output is

$$\frac{\hat{y}}{\hat{x}} = \frac{\mu}{1 + \mu B} = \frac{k_1 k_2 k_3}{as^2 + bs + c + k_1 k_2 k_3 k_4} \quad (11.66)$$

where μ is the forward transfer function and B is the transfer function through the feedback path of the loop. The transfer function \hat{y}/\hat{x} thus has a simple quadratic denominator, assuming that the time constants of the several spring-mass systems are very small. In the FC2 valve the loop-gain constant $k_1 k_2 k_3 k_4$ is about 314, from data furnished by the manufacturer, and the response of the system is as shown in Fig. 11.46.

The frequency responses of the Moog and Cadillac valves are similar with the exception that the FC2 valve amplitude response falls off more gradually than the typical Moog characteristic. However, the damping

factor of both valves may be adjusted within a limited range by changing internal dimensions.

In all the hydraulic valves so far discussed, including the flapper valves, the transfer is from the input variable to flow. Constant flow will cause a constant velocity in the load element. If the over-all control system controls the position of this element, the loop contains one integration.

Occasionally, an extremely high performance control system is required, and the designer must provide two integrations within the loop to meet dynamic accuracy specifications. Both the Moog and Cadillac valves can be adapted to provide two integrations. In the Moog valve the centering spring may be removed, while in the Cadillac valve the feedback spring and lever may be removed. The force set up by the differential pressure in the chamber will then move the piston at a constant velocity, and the oil flow will constantly increase. Naturally, such a device must be part of a feedback loop to operate properly, and the loop must be properly equalized or compensated if it is to be stable.

In certain applications both the Moog Model 500 valve and the Cadillac FC2 valve encounter problems with magnetic dirt building up around the poles of the torque motor and hindering proper operation. Both concerns have developed dry torque-motor designs to overcome this difficulty.

11.13. Pressure-regulating Devices. Several different methods of obtaining constant-pressure hydraulic supplies are used, depending on

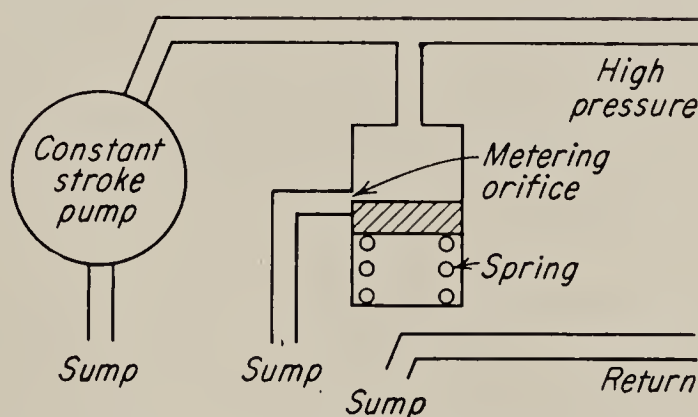


FIG. 11.47. Elementary form of pressure regulator. Constant-flow pump with bypass orifice for constant-pressure service.

the type of pump and the accuracy with which the pressure must be controlled.

One type of pressure control is shown in simplified form in Fig. 11.47. It consists of a spring-loaded valve connected from the high-pressure line to sump. This type of pressure control is used with fixed-stroke pumps. The flow from the pump is essentially constant at full flow, with the valve porting to sump any oil in excess of that required by the load. There is a considerable waste of flow and power in the regulating valve unless the system is always operating at close to full load. The wasted power results in a temperature rise in the hydraulic fluid and may require

special provisions for cooling. Commercial devices operating on this principle may be more elaborate.

A second method of pressure control consists of a pressure-sensitive actuator and a variable-stroke pump, as shown in Fig. 11.48. The main advantage of this method is the saving in power and flow. The pump supplies only that flow required by the load. The major disadvantage of this device is the extra expense of the variable-stroke pump.

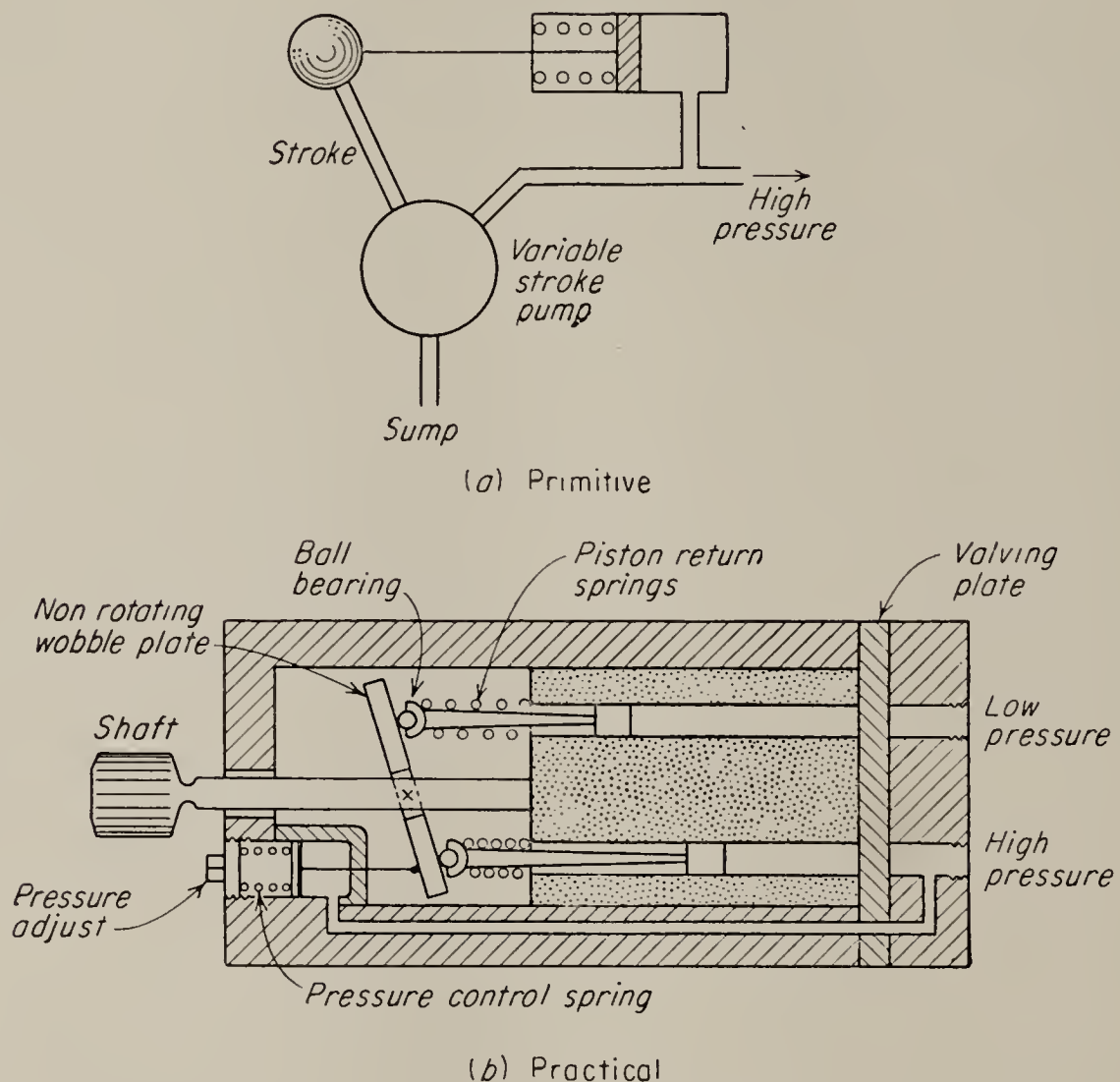


FIG. 11.48. Variable-stroke pump with pressure-controlled stroke for constant-pressure service.

In constant-pressure supply systems a storage reservoir, or *accumulator*, is usually included to smooth variations in pressure due to pump ripple and to supply the system during periods of peak demand. In the case of intermittent operation of actuators, the hydraulic pump may be completely inadequate to supply peak demand, but it serves to charge the accumulator during periods of inaction.

The accumulator may be of the *bellows type*, in which the high-pressure fluid expands a bellows, or it may be in the form of a cylinder in which the pressure actuates a piston that compresses a spring. Repeated cycling of the bellows type causes fatigue, and hence this type is unsatisfactory in general. The most satisfactory type of accumulator for high-performance systems is the *hydropneumatic type* in which the spring action is obtained from compression of a gas. The inflexible-separator or piston

type is generally not so desirable as the newer flexible-separator type. Owing to the inertia of the piston the pressure peaks may actually be increased in the piston type. The flexible-separator hydropneumatic accumulator is the better design. It yields more power output and is more reliable for a minimum weight and size.¹

11.14. Choice of Operating Pressure for Hydraulic Systems. One of the first decisions to be made by the designer of a flow-control hydraulic servomechanism is the choice of the system operating pressure. Early hydraulic mechanisms were usually operated in the pressure range of several hundred pounds per square inch. The operating pressure was chosen empirically and gradually increased as experience was gained in the field. Modern aircraft hydraulic systems have been designed for pressures as high as 4,000 to 5,000 psi.

Several factors enter into the choice of system pressure. As the operating pressure for a given load is increased, the following effects are noted:

1. The required flow is decreased, thus allowing the use of smaller displacement pumps, valves, and actuators, as well as smaller lines.

2. The weight of the fluid required decreases continuously, and the weight of the components decreases up to a value of about 4,000 psi,* at which point the strength of the material dictates an increase in size and weight.

3. The operating temperature of the system steadily increases as a result of the increased leakage at valves and seals.

4. The danger of fire and explosion at extremely high pressures dictates the use of nonflammable artificial hydraulic fluid, which is more expensive by a factor of 10 than petroleum-base products.

5. At the very high pressures manufacturing difficulties prevent taking full advantage of the reduction in size of valves and actuators. All of these factors have influenced the choice of 800 to 1,500 psi as standard in most modern high-performance control systems. In the United States, 3,000 psi is standard in aircraft applications, because weight reduction is of prime importance and the expense factor is secondary, whereas certain British aircraft such as the Bristol Britannia operate at 4,000 psi. Proposals for raising the standard pressure in U.S. aircraft systems from 3,000 to 4,000 psi will result in an improvement of only 2½ per cent in the figure of merit and are thus not worth while, according to Cooke.²

PROBLEMS

11.1. Show from Eq. (11.5) the result of making a in Fig. 11.1 equal zero.

11.2. In Fig. 11.49 is shown a typical spring-centered spool valve. Find the transfer function \hat{x}_o/\hat{F}_L for operation about the point $x_v = 0$, $p_L = 0$ for the hydraulic-

¹ E. M. Greer, Hydraulic Accumulators, *Machine Design*, vol. 26, no. 1, p. 132, 1954.

* C. Cooke, Optimum Pressure for a Hydraulic System, *Product Eng.*, vol. 27, no. 5, p. 162, 1956.

² *Ibid.*

actuator arrangement shown. The valve is completely symmetrical and slightly underlapped; hence when the valve piston is at the exact center ($x_v = 0$), each orifice is open. The amount of underlap is such that a motion of the valve piston ± 0.001

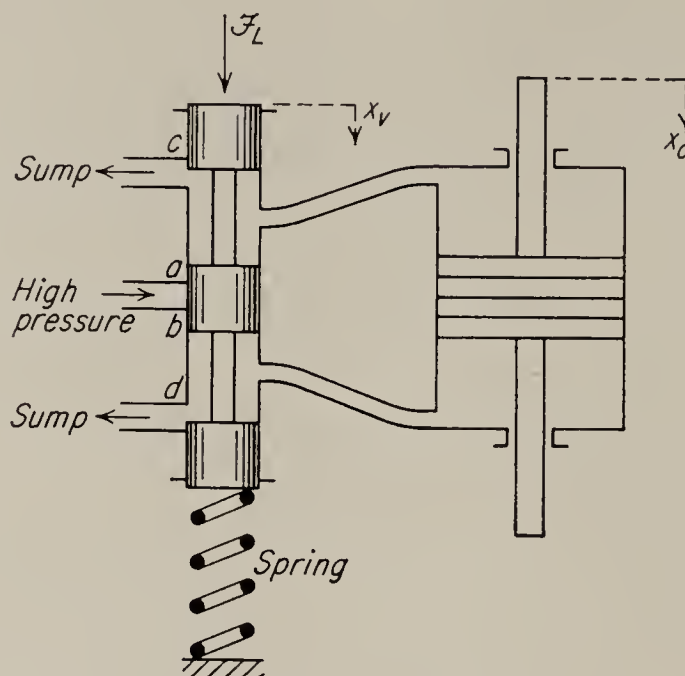


FIG. 11.49

in. from the neutral position will just close one set of orifices. Assume that the orifices do not leak at all when they are closed and produce no steady-state hydraulic reaction force. Also assume that transient hydraulic forces are negligible.

Each orifice is characterized by the two equations

$$F_i = Cx_i \sqrt{\frac{2g}{\rho}} p_i \quad \text{and} \quad \mathfrak{F}_i = kx_i p_i$$

where $C, k = \text{constants}$

$g = \text{acceleration of gravity}$

$\rho = \text{density of oil} = 0.03 \text{ lb/in.}^3$

$x_i = \text{amount of opening of } i\text{th orifice, in.}$

$F_i = \text{flow through } i\text{th orifice, in.}^3/\text{sec}$

$p_i = \text{pressure drop across } i\text{th orifice, psi}$

$\mathfrak{F}_i = \text{force of hydraulic reaction produced at } i\text{th orifice, lb}$

The following data are given:

Mass of valve piston = 0.1 lb

Valve spring constant = 200 lb/in.

Valve coefficient of viscous friction = 0.04 lb-sec/in.

Mass of actuator piston = 1.5 lb

Area of actuator piston = 1 in.²

Friction of actuator piston = 1 lb-sec/in.

When $x_v = 0.005$ in., when supply pressure = 2,000 psi, and when there is no load on actuator, the piston moves at a rate of 5 in./sec and the total hydraulic reaction force is 0.9 lb.

11.3. Assume that at each of the valving orifices shown in Fig. 11.49 the relation between pressure flow and displacement is

$$F_i = \frac{Cx_i}{1 - e^{\alpha x_i}} \sqrt{\frac{2\rho}{g}} p_i$$

where F_i = flow through i th orifice, in.³/sec

C, α = constants

ρ = density of oil = 0.03 lb/in.³

g = acceleration of gravity

p_i = pressure drop across i th orifices, psi

x_i = spool displacement at each orifice, in.

Note that, for orifices a and c , $x_i = x_v$, and for orifices b and d , $x_i = -x_v$. When the valve is centered, $x_v = x_i = 0$. Each orifice produces a hydraulic reaction force given by the relation

$$\mathfrak{F}_i = KF_i \sqrt{p_i}$$

All the forces \mathfrak{F}_i are in such a direction as to tend to close the valve.

The following data are given:

Mass of valve piston = 0.1 lb

Valve spring constant = 200 lb/in.

Valve coefficient of viscous friction = 0.04 lb-sec/in.

Mass of actuator piston = 1.5 lb

Area of actuator piston = 1 in.²

Friction of actuator piston = 1 lb-sec/in.

When $x_v = 0.005$ in., when supply pressure = 2,000 psi, and when there is no load on actuator, the piston moves at a rate of 5 in./sec and the total hydraulic reaction force is 0.9 lb. When $x_v = 0$, when supply pressure = 2,000 psi, and when a force of 1,000 lb is applied to the piston, the piston moves at a rate of 0.01 in./sec.

Find the transfer function $\hat{x}_o/\hat{\mathfrak{F}}_i$ for operation about the point $x_v = 0$, $p_L = 0$.

11.4. Find the axial force developed at one port of a spool valve with a piston diameter of $\frac{1}{4}$ in. and a rectangular port of 0.1-radian width. Assume the stroke is $\frac{1}{16}$ in. and the radial clearance is 0.002 in. Assume an orifice coefficient of 0.6 and an operating pressure of 3,000 psi.

11.5. The flow relation of a hydraulic valve is given by

$$F = 25x \sqrt{p_v}$$

where F = flow, in.³/sec

x = valve-spool displacement, in.

p_v = total pressure drop across valve, psi

The valve is connected to a constant-pressure supply of 1,500 psi. The sump pressure is zero. A hydraulic motor with the following characteristics is connected to the servo ports of the valve:

Moment of inertia = 0.003 lb-in.-sec²

Viscous friction = 0.2 lb-in.-sec

Hydraulic displacement = 0.38 in.³/rev

a. Find the response of motor speed versus time when x is given a step displacement from zero to 0.01 in.

b. Find the response of motor speed versus time when x varies sinusoidally at a frequency of 10 cps with an amplitude of 0.01 in.

11.6. Calculate the orifice diameters for a flapper valve to operate at 4,000 psi. Allow a maximum flow of 0.05 gpm through the nozzle. Calculate the reaction force on the flapper when it is in the center of its operating range. The density of the oil is 0.03 lb/in.³.

11.7. Assume that the nozzle in Fig. 11.32 has been imperfectly made and that the jet emerges at an angle of 10° from the center line of the jet pipe at 1,500-psi supply pressure and an orifice diameter of 0.1 in. What is the effect?

11.8. A flapper valve operating at 1,500-psi supply pressure has an upstream orifice with a diameter of 0.005 in. and a controlled-orifice diameter of 0.012 in. (see Fig. 11.50). The piston has a large area of 0.05 in.² and a small area of 0.025 in.². The piston mass is 0.02 lb and the viscous friction between the piston and cylinder is such that a force of 0.04 lb is required to move the piston at a rate of 1 in./sec. Obtain a small-signal transfer function between flapper position and piston position for the operating point defined by the fact that the piston is stationary in the middle of its allowed travel, and there is no external force on it. Assume that the coefficient of discharge for

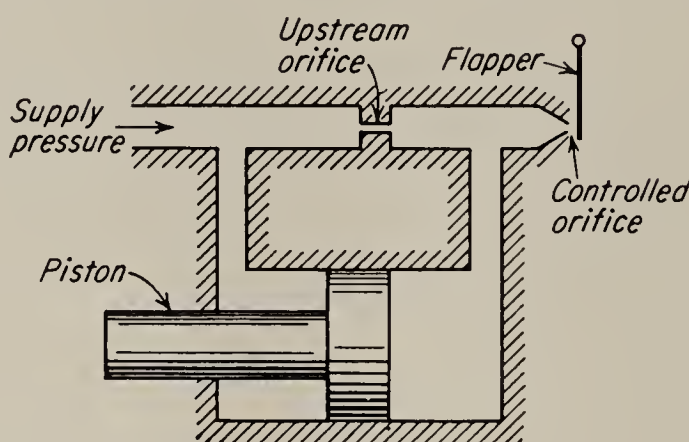


FIG. 11.50

both orifices is 0.65 and that the compressibility of the fluid in the chamber is negligible. The density of the fluid is 0.03 lb/in.³

11.9. The flapper of the valve described in Prob. 11.8 is spring-restrained such that the hydraulic reaction force at the quiescent operating point is balanced. The spring constant of the spring used for this purpose is 100 lb/in. The flapper is controlled by an electric torque motor which supplies a force of 0.05 lb/ma. The mass of the flapper is negligible. Find the small-signal transfer function from input current to output position of the piston for this system. Consider operation around the operating point defined in Prob. 11.8.

11.10. A valve of the Moog type (see Fig. 11.41) has a second-stage valve piston with a diameter of 0.1875 in. and a weight of 0.02 lb. The spring constant of each one of the springs used to center the piston is 400 lb/in. The control orifices for the inlet and sump are rectangular in shape and 0.03 in. wide. There are four of these rectangular holes spaced equally around the circumference of the sleeve at each land of the valve piston. The first stage is a flapper valve having fixed orifices of 0.005-in. diameter and variable orifices of 0.012-in. diameter. The flapper is controlled by a torque motor supplying a force at the flapper of 10 lb/amp. The flapper is spring-restrained, but has negligible mass or damping. The chamber volume is small enough so that compressibility of the oil can be neglected.

The flow gain of this valve is measured by connecting the two outlet ports together with a pipe large enough so as to produce negligible pressure drop. A flow meter is then placed into the return line and the valve is connected to a 1,500-psi supply. After subtracting the first-stage leakage flow, the flow gain is found to be 2-in.³/sec/ma current difference into the torque motor.

Using small-signal methods about the operating point defined by load pressure = load flow = 0, supply pressure = 1,500 psi, find the transfer function of the valve. The density of the fluid is 0.03 lb/in.³

11.11. For the slide valve shown in Fig. 11.33, show that the reaction force tends to open the valve when the slider is displaced from dead center.

CHAPTER 12

PNEUMATIC SYSTEMS

12.1. Introduction. Control systems that transmit power by means of fluid flow are subdivided into two classifications. Hydraulic systems constitute the first class and have been considered in the preceding chapters. The second class comprises pneumatic systems. The division is made on the basis of the importance of compressibility of the flow medium. Roughly, the division can be made on the state of the fluid. The medium for hydraulic systems is a liquid, while for pneumatic systems the fluid is a gas.

In general, a pneumatic control is “softer” than a hydraulic system. The torque of a pneumatic motor builds up slowly with respect to time, and the system will yield under shock loads. While this contributes to the long life of the components, compressibility complicates the stability problem in closed-loop systems. In pneumatic systems there is no possibility of harmful shock waves when the flow is suddenly stopped; this problem does exist in hydraulic systems and is sometimes called the “water-hammer” effect.

Pneumatic systems have certain other advantages over hydraulic systems. First, there is no fire hazard. Although artificial hydraulic fluids that are nonflammable have been developed, petroleum-base fluids are still preferred for hydraulic systems because of their anticorrosion and lubricating qualities as well as their lower cost. Even minute hydraulic leaks in high-pressure systems can fill the surrounding atmosphere with atomized hydraulic vapor, thus creating an extreme fire and explosion hazard.

Second, since, in a pneumatic system, the air can be vented to the atmosphere at the actuator, it is often possible to design the system with only one line, while in a hydraulic system this is not possible. This cuts down on the cost of fittings and lines, etc. It also results in a weight reduction, which is the third advantage.

The weight of pneumatic lines and fittings is less than the weight of equivalent hydraulic fittings. There is also the further reduction in weight due to the elimination of the return line, and, finally, the weight of the hydraulic fluid itself is eliminated in pneumatic systems. In an

interesting comparison between pneumatic and hydraulic systems for the Convair Model 240 aircraft, Gerwing and Famme¹ found that the pneumatic system weighed exactly one-half as much as the hydraulic system. A more detailed weight comparison of electric, hydraulic, and pneumatic components is given in the following chapter.

A fourth advantage of pneumatic systems is the ease of maintenance. When components must be replaced in a hydraulic system, the fluid must be drained before replacement can proceed. Furthermore, fluid "weepage," or the oil caught behind seals and connections during assembly that oozes out over a period of time, must be cleaned off periodically in hydraulic systems.

A fifth advantage that can be important if the control system is subjected to extremes in temperature is that the viscosity of the fluid in pneumatic systems is usually negligible. Furthermore the variation in viscosity with temperature is small; thus, even when viscosity is considered, the effect of temperature will be small. This is not so in hydraulic systems. There viscosity is a first-order effect, and the variation of viscosity with temperature is a major problem.

There are several disadvantages of pneumatic systems as compared with hydraulic systems. First, the lubrication of components must be carefully considered, since this is not automatically taken care of as in hydraulic systems. Second, more work is required to pump a compressible fluid up to a given pressure than is required by a noncompressible fluid. This work must be charged to the efficiency of the system if the work is not recovered by the actuator. The third disadvantage of a pneumatic system is the possibility of explosion if the pneumatic storage tank, or accumulator, is punctured. It will be seen in Chap. 13 that, in order for the maximum weight savings to be realized with a pneumatic system, a large accumulator must be provided. The large amount of energy that is stored in the compressed gas will be explosively released if the tank is ruptured. A break in a hydraulic system is messy and a potential fire hazard, but the same danger of explosive expansion does not exist. Modern pneumatic-component design attempts to render the accumulator penetration-proof.

Finally, the compressibility of the fluid adds equivalent springs to the system in various places, and thus lags and time constants are developed. This is a fundamental limitation on the high-speed response of a pneumatic system. In addition to the lags added to the system by compressibility, the decreased density of the fluid results in a slower signal-transmission rate. A line length that would be considered negligible in a hydraulic system could have a major effect on the operation of a pneu-

¹ H. F. Gerwing and J. H. Famme, Pneumatic versus Hydraulic Systems for Aircraft, *Product Eng.*, vol. 21, no. 11, p. 21, 1950.

matic system. Basically a pneumatic system is slower than an equivalent hydraulic system. Shearer and Lee¹ compare a simple hydraulic positioning system with an equivalent pneumatic system operating at the same pressure and find the hydraulic system about forty-five times faster. This disparity could be reduced, however, by operating the pneumatic system at a considerably higher pressure while still not exceeding the limits of safety.

Probably the oldest application of pneumatic systems is in low-pressure process control in the chemical industry. For many years process-control engineers were alone in their attempts to analyze closed-loop pneumatic systems, and understandably they have built up a jargon of their own. Unfortunately their nomenclature differs considerably from that used in the newer and more widely known control fields. There is evidence that this lack of communication has prevented the use by the rest of the field of techniques developed many years ago by process-control engineers and has hindered progress in the process-control industry more recently by hampering the use of work done outside that field.

12.2. Pneumatic Flow. Usually in the analysis of a pneumatic system the viscosity of the medium is ignored, since the viscosity of air is about 10^{-3} that of hydraulic fluids. While the analysis of the flow relations proceeds in exactly the same way as the previous analysis of incompressible fluids, different factors are important. The general energy equation from which the Bernoulli equation was developed is

$$q + \frac{p_1 v_1}{778} - \frac{p_2 v_2}{778} + \frac{\text{work}}{778} = u_2 - u_1 + \frac{V_2^2 - V_1^2}{2g(778)} + \frac{z_2 - z_1}{778} \quad (12.1)$$

where q is the heat transferred to the fluid, p is the pressure in pounds per square foot, work is the mechanical work done by the fluid, u is the internal energy, V is the velocity in feet per second, z is the elevation, and v is the specific volume, or volume per unit weight. Usually no heat is transferred to the fluid, and thus q is zero. A process for which q is zero is called an *adiabatic process*. Furthermore the difference in head or elevation is negligible, and in a flow process no work is done by the fluid. Then

$$\frac{p_1 v_1}{778} - \frac{p_2 v_2}{778} = u_2 - u_1 + \frac{V_2^2 - V_1^2}{2g(778)} \quad (12.2)$$

This equation applies whether the fluid is liquid, gas, or vapor. We shall assume that our fluid follows the perfect-gas law $pv = RT$, where R is the gas constant and T is the absolute temperature. For air, R is 53.3 ft/°R. It can be shown² that, for any perfect-gas process, the change in internal

¹ S. L. Shearer and S. Y. Lee, Selecting Power Control Valves, *Control Eng.*, vol. 3, no. 4, pp. 73–76, 1956.

² Binder, "Fluid Mechanics," Prentice-Hall, Inc., Englewood Cliffs, N.J., 1949, p. 197.

energy is

$$u_2 - u_1 = c_v(T_2 - T_1) \quad (12.3)$$

where c_v is the specific heat of the fluid at constant volume. It can be further shown¹ that

$$c_v = \frac{1}{778} \frac{R}{k - 1} \quad (12.4)$$

For a general polytropic process

$$pv^k = \text{constant} \quad (12.5)$$

and for air, $k = 1.4$, if the process is adiabatic, and $k = 1.0$, if the process is isothermal. Thus by combining Eqs. (12.3) to (12.5), Eq. (12.2) may be written as

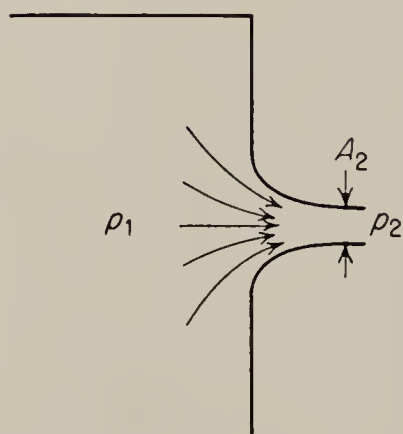
$$\frac{V_2^2 - V_1^2}{2} = \frac{k}{k - 1} \frac{p_1}{m_1} \left[1 - \left(\frac{p_2}{p_1} \right)^{(k-1)/k} \right] \quad (12.6)$$

where m_1 is the initial mass density of the fluid. Thus by definition,

$$m_1 = \frac{w_1}{g} = \frac{1}{v_1 g} \quad (12.7)$$

where w_1 is the initial weight density and v_1 the initial specific volume.

12.3. Pneumatic Flow through an Orifice. Let us consider the case of adiabatic flow through a nozzle or orifice, as shown in Fig. 12.1. We shall assume that V_1 , the velocity on the high-pressure side, is negligible with respect to V_2 , the velocity at the outlet. Thus Eq. (12.6) becomes



$$V_2 = \sqrt{\frac{2k}{k - 1} \frac{p_1}{m_1} \left[1 - \left(\frac{p_2}{p_1} \right)^{(k-1)/k} \right]} \quad (12.8)$$

Let W be the weight of gas flowing per second and A_2 be the area at the throat. Then

$$W = \frac{A_2 V_2}{v_2} \quad (12.9)$$

where v_2 is the specific volume of the gas at the throat of the nozzle. Substituting for V_2 in Eq. (12.9) and employing Eqs. (12.5) and (12.7) results in

$$W = A_2 \sqrt{\left(\frac{2kg}{k - 1} \right) \frac{p_1}{v_1} \left[\left(\frac{p_2}{p_1} \right)^{2/k} - \left(\frac{p_2}{p_1} \right)^{(k+1)/k} \right]} \quad (12.10)$$

Figure 12.2 shows a plot of Eq. (12.10), the weight flow through an orifice as a function of p_1/p_2 . The maximum flow occurs at a ratio $p_2/p_1 = 0.53$.

¹ *Ibid.*

This may be found from Eq. (12.10) by differentiating. For pressure drops greater than the critical drop, Eq. (12.10) shows the flow decreasing. Actually this does not occur; rather the flow remains constant, no matter how low the downstream pressure becomes. This is an effect similar to the venturi tube effect discussed in Chap. 11. The velocity of flow becomes supersonic above the critical pressure ratio, and no downstream condition can influence upstream conditions.

If in Eq. (12.10) we let $p_2/p_1 = 0.53$ and $k = 1.4$, the relation for flow through a converging orifice at greater than critical pressures may be found to be

$$W = 0.7A_2 \sqrt{g \frac{p_1}{v_1}} = 0.7A_2 \sqrt{\frac{gp_1}{RT/p_1}} \quad (12.11)$$

For air, $R = 53.3$; thus

$$W = \frac{0.53A_2 p_1}{\sqrt{T_1}} \quad (12.12)$$

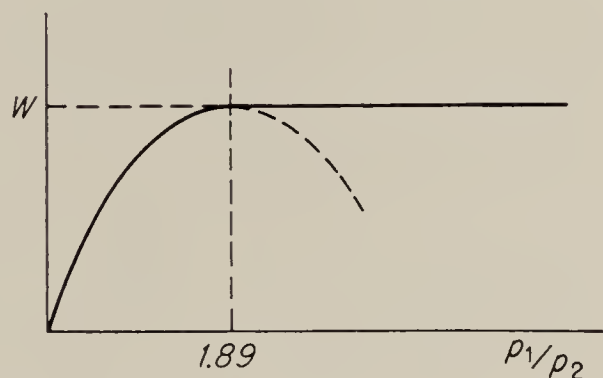


FIG. 12.2. Weight flow through orifice.

where T_1 is the absolute temperature in degrees Rankine. Equation (12.12) is sometimes called Fliegner's equation. Thus for pneumatic flow, but not for hydraulic flow, there is a linear relation between upstream pressure and flow in the region above critical pressure drop, and in this region flow is independent of downstream pressure. Since this relationship holds so long as the pressure drop across the orifice is more than 1.89 to 1, pneumatic systems in general are designed to operate in this region to take advantage of this fact. Just as in hydraulic flow, the actual measured area of the orifice is not the area of the jet flow, nor does this relation include the effects of turbulence, viscosity, or converging flow. These effects reduce the mass rate of flow through the orifice and are usually accounted for by an orifice coefficient of between 0.6 and unity. The Compressed Air Institute recommends 0.65 for sharp-edged orifices and 0.77 for well-rounded entrances. Stenning¹ reports that values between 0.8 and unity apply for servo-valve-type orifices.

12.4. Compressible Flow in Pipes. The general equation for compressible flow in pipes may be derived in the same manner as the equation for hydraulic flow, which was obtained in Chap. 10. The general differential equation for compressible flow is²

$$-dp \frac{\pi D^2}{4} - \frac{fV^2}{8gv} \pi D dl = \frac{\pi D^2}{4} \frac{dl}{gv} V \frac{dV}{dl} \quad (12.13)$$

¹ A. H. Stenning, An Experimental Study of Two-dimensional Gas Flow through Valve-type Orifices, ASME paper 54-A-45.

² See, for instance, Binder, *op. cit.*, p. 229.

where f is the pipe friction factor, which can be expressed as a function of the Reynolds number and can be determined from the Staunton diagram in Fig. 10.5. Rearranging (12.13), we obtain

$$v dp + \frac{V dV}{g} + \frac{fV^2 dl}{2g D} = 0 \quad (12.14)$$

and

$$\frac{2gv}{V^2} dp + 2 \frac{dV}{V} + \frac{f dl}{D} = 0 \quad (12.15)$$

Usually certain simplifying assumptions concerning the nature of the flow may be made before integration is attempted. One type of flow is isothermal, or constant-temperature, flow; the second type of flow is adiabatic, or no-heat-loss, flow. Actual pneumatic processes lie between these two extremes. Experimental work on many types of pneumatic components such as flapper valves, capillary tubes, and tapered-pin orifices have shown excellent agreement with theoretical isothermal flow.¹ For isothermal flow $pv = RT$, and Eq. (12.15) can be integrated as

$$p_1^2 - p_2^2 = \frac{V_1^2 p_1}{g v_1} \left(2 \ln \frac{V_2}{V_1} + \frac{fl}{D} \right) \quad (12.16)$$

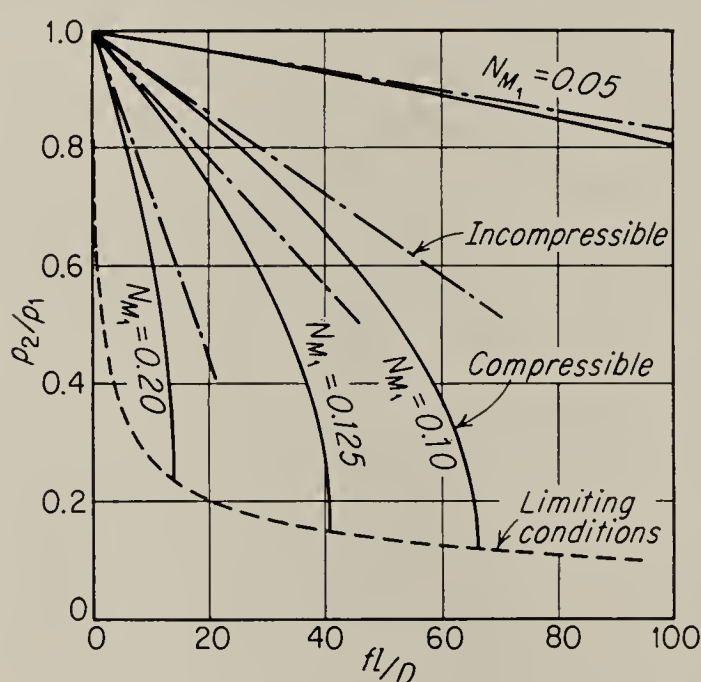


FIG. 12.3. Pressure versus pressure ratio for various Mach numbers.

Figure 12.3 shows normalized pressure versus the pressure ratio fl/D for various initial Mach numbers, where the Mach number is defined as

$$N_{m_1} = \frac{V_1}{\sqrt{gkRT_1}} \quad (12.17)$$

and in this case we take $k = 1.4$ for air. For comparison, the plot for incompressible fluids under the same conditions is also given in Fig.

12.3. It may be seen that, for low Mach numbers and short lengths of pipe, incompressible flow may be assumed with little error.

Quite often the logarithmic term in Eq. (12.16) can be neglected with negligible error. Under this assumption (12.16) becomes

$$p_1^2 - p_2^2 = \frac{V_1^2 p_1 fl}{g v_1 D} \quad (12.18)$$

For Reynolds numbers less than 2,000 the flow is usually laminar, and f may be represented by the empirical relation found from the Staunton

¹ C. R. Webb, Nonlinearities in a Pneumatic Controller, *Trans. Soc. Instrument Tech.*, vol. 7, no. 1, p. 9, 1955.

diagram of

$$f = \frac{64}{N_R} \quad (12.19)$$

Substituting the relation for volumetric flow for V_1 and rearranging, there results

$$F = \frac{\pi(p_1^2 - p_2^2)D^4}{128\mu l 2p_1} \quad (12.20)$$

Equation (12.20) has been arranged for comparison with Eq. (10.22), the Hagen-Poiseuille law for incompressible laminar flow in a pipe. While there are some similarities, the relations are not the same. Figure 12.3 shows the error involved in using the incompressible relations instead of the compressible relations for various Mach numbers.

12.5. Compressibility as an Equivalent Spring. Even if compressibility is ignored in the flow through pipes, it should be considered for its effect on the transfer function of the various system components. Just as a viscous fluid adds viscous damping to various components, a compressible fluid acts as an equivalent spring. For instance, the resonant frequency of a typical spool-type pilot valve was increased from 700 radians/sec to 950 radians/sec when 1,000-psi air was introduced into the system; this result was due to the equivalent spring action of the air supply. Usually the viscous-damping effect of compressed air is negligible compared with friction and other sources of damping.

The spring rates of trapped compressed air can be important in actuator cylinders. Figure 12.4* gives the equivalent spring constant of compressed air initially at 100 psi trapped behind the piston in cylinders of various sizes. The relation for the equivalent spring constant is

$$k = \frac{\pi D^2 p_0}{L} \quad (12.21)$$

where k = spring rate, lb/in.

D = cylinder diameter, in.

L = cylinder length, in.

p_0 = initial pressure behind piston, psi

12.6. Pneumatic Transmission Lag. Another effect that should be considered in pneumatic systems is the transmission- or transport-lag effect. Since air is much less dense than hydraulic fluid, the velocity of propagation is lower. The approximate figure for the speed of sound in air may be taken as 1,100 ft/sec, about one-fifth the speed of sound in hydraulic oil. The actual signal-propagation velocity (group velocity) will be less than this. If the pneumatic line were a perfect transmission

* H. Levenstein, Pneumatic Servomechanisms, *Control Eng.*, vol. 2, no. 6, p. 67, 1955.

line, the output would reproduce perfectly any inputs, delayed in time, however. The Laplace transfer function for a perfect transmission line may be found by employing the real translation theorem.¹ If $\mathcal{L}[f(t)] = F(s)$, it can be shown that

$$\mathcal{L}[f(t - t_0)] = F(s)e^{-t_0s} \quad (12.22)$$

Unfortunately actual pneumatic transmission lines are not perfect; they attenuate and distort the signal. In Fig. 12.5 is given the experimental

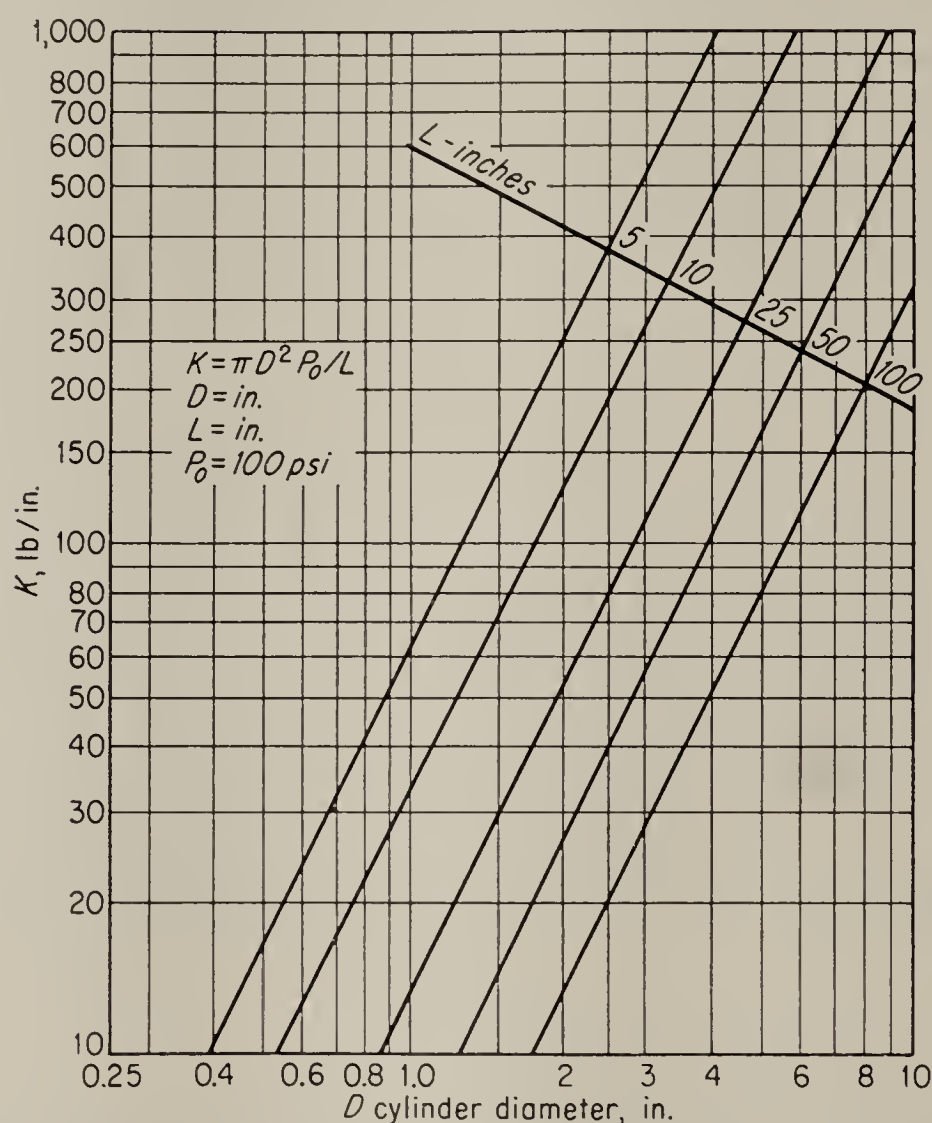


FIG. 12.4. Equivalent spring rates of compressed air. Trapped air was initially at 100 psi.

response of a long pneumatic line to a step-function input.² The delay time appears to be about 3.0 sec, which is more than twice the time that would be calculated if the theoretical velocity of sound in air were employed. Furthermore, there is a high-frequency attenuation that rounds off the sharp edge of the step function.

As a first-order approximation, a transmission line might be represented

¹ Gardner and Barnes, "Transients in Linear Systems," John Wiley & Sons, Inc., New York, 1950, p. 236.

² M. Bradner, Pneumatic Transmission Lag, *Instruments*, vol. 22, p. 618, 1949.

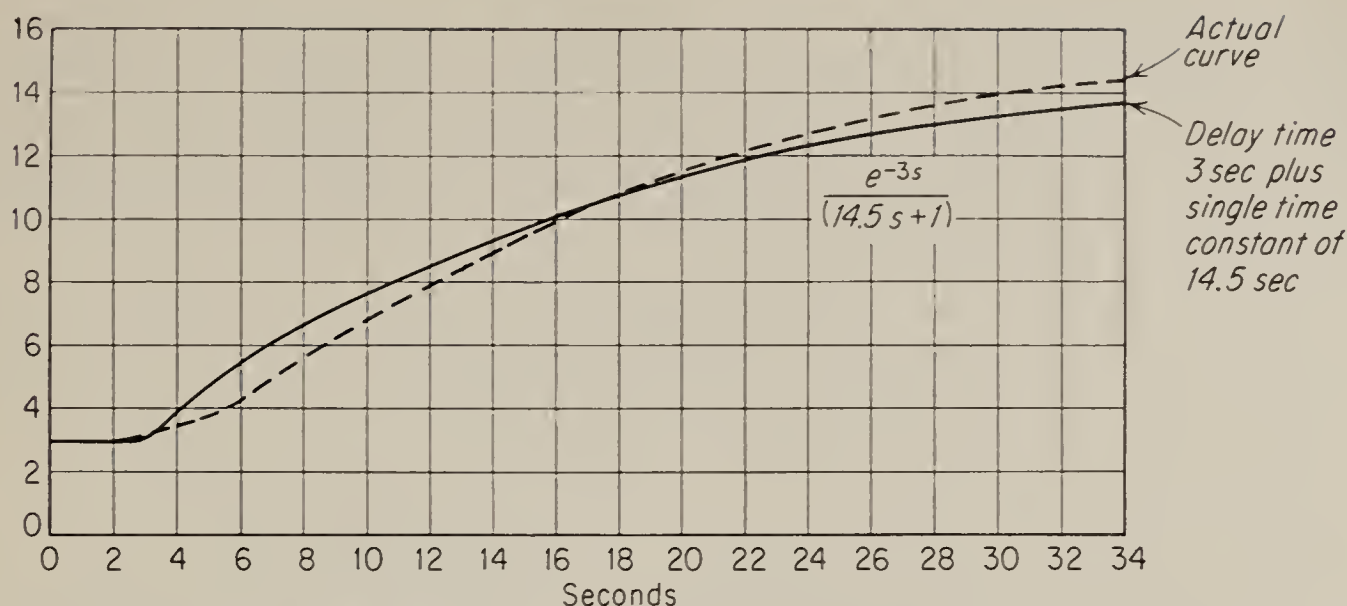


FIG. 12.5. Experimental response of a long pneumatic line to a step-function input.

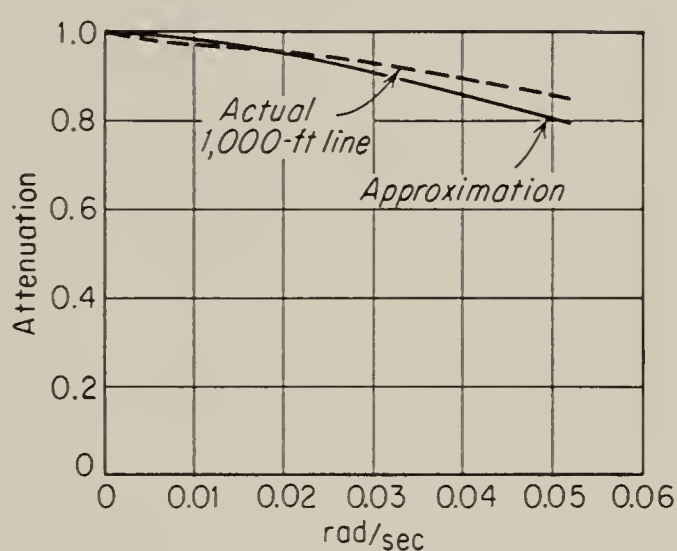
by a delay time and a single time constant, or

$$\frac{\hat{y}}{\hat{x}} = \frac{e^{-t_0 s}}{1 + Ts} \quad (12.23)$$

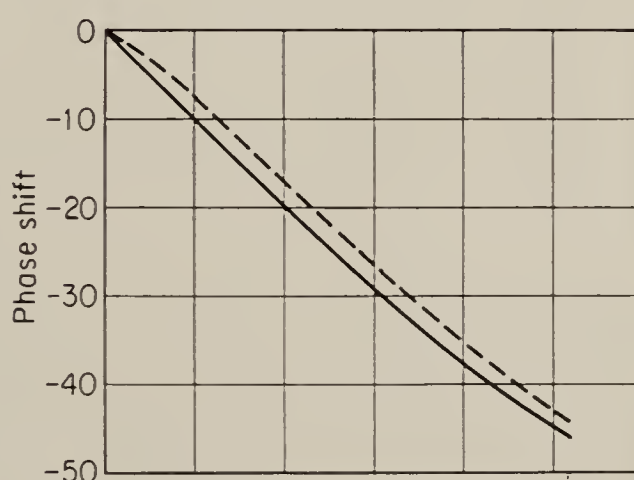
where t_0 is the delay time and T is the time constant. Shown in Fig. 12.5 for comparison with the actual response of the line is the response of this approximation to a step function of applied pressure. t_0 is taken as 3 sec, and T was taken as 14.5 sec. The time constant T was taken from the actual curve at 63 per cent of the final value. In Fig. 12.6 are shown the frequency response of the actual line, as obtained by Bradner, and the response of the approximation for comparison. It may be seen from Figs. 12.5 and 12.6 that the approximation compares quite favorably with the actual test data. For more accurate results the exact transmission-line equations given in Chap. 10 should be employed.

12.7. Pneumatic-Electric Analogues and Pneumatic Control Functions.

Just as it was possible to find electric analogues for mechanical and hydraulic elements, it is possible for pneumatic elements. Usually pres-



(a) The attenuation



(b) The phase shift

FIG. 12.6. Frequency response of a pneumatic transmission line. This is the line whose response is shown in Fig. 12.5, and the single-time-constant approximation is also shown here for comparison.

sure is represented as voltage, and volumetric flow as current. Stolarik¹ has developed resistance elements consisting of baffle plates and capillary tubes that are essentially linear in their behavior. Capacitors are tanks or reservoirs. Figure 12.7 shows a pneumatic resistance and capacitance network and its electric equivalent.

Using the pneumatic equivalents of resistance and capacitance, it is possible to construct pneumatic networks to given frequency-response specifications, thus making possible the synthesis of stabilizing networks for closed-loop pneumatic systems.

The particular network to be constructed depends on the control function that is to be synthesized. In Chap. 1 is given a list of control functions and the electric networks that generate them. The pneumatic device may be constructed by analogy to these electric networks. The magnitudes of the

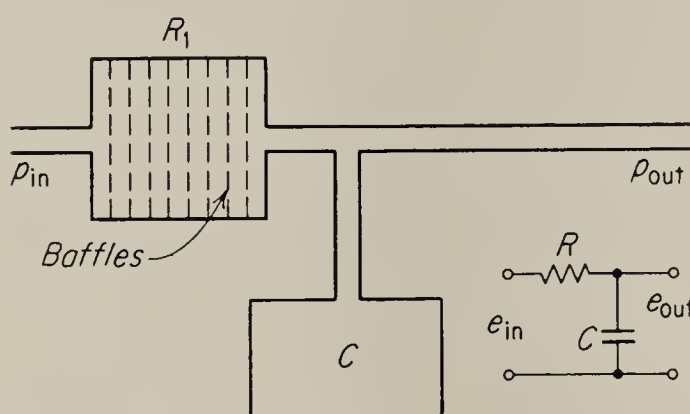


FIG. 12.7. A pneumatic high-pass network and its electric equivalent. (After Stolarik)

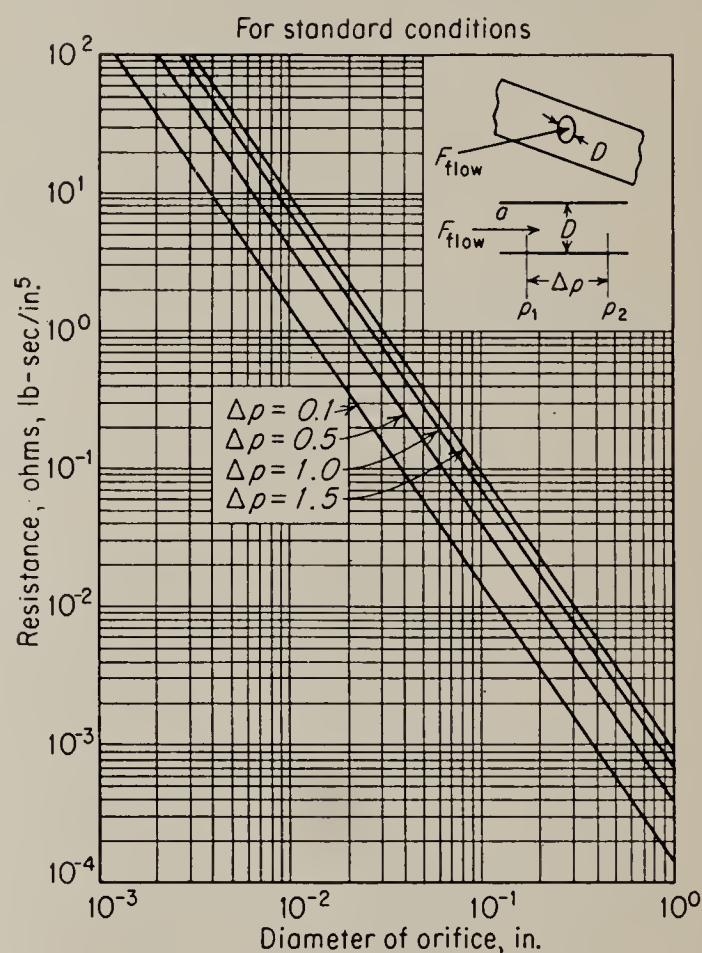


FIG. 12.8. Resistance of a round orifice. (Stolarik)

various pneumatic components may be determined by calculation or by the use of Figs. 12.8 to 12.13. These curves are also valuable in calculating the unavoidable lags due to given pneumatic components.

Figures 12.8 through 12.11 relate pneumatic resistance to variously shaped constrictions. Pneumatic resistance is defined by analogy to electric resistance as

$$R = \frac{\Delta P}{F} \quad \text{lb-sec./in.}^5 \quad (12.24)$$

where ΔP is the pressure drop and F is the volumetric flow. The pressure-flow relations have been derived in Secs. 12.3 and 12.4. These relations

¹ E. Stolarik, Notes on Pneumatics, *Aero Dig.*, vol. 60, no. 1, p. 45, no. 2, p. 44, 1950.

are merely presented in a convenient form in the following curves. The resistance is derived for air at standard conditions of 15°C and 14.7 psi and for an orifice discharge coefficient of 0.62. While the resistance of capillary tubes is essentially independent of pressure, it does depend on temperature, and the resistance of orifices depends on both temperature

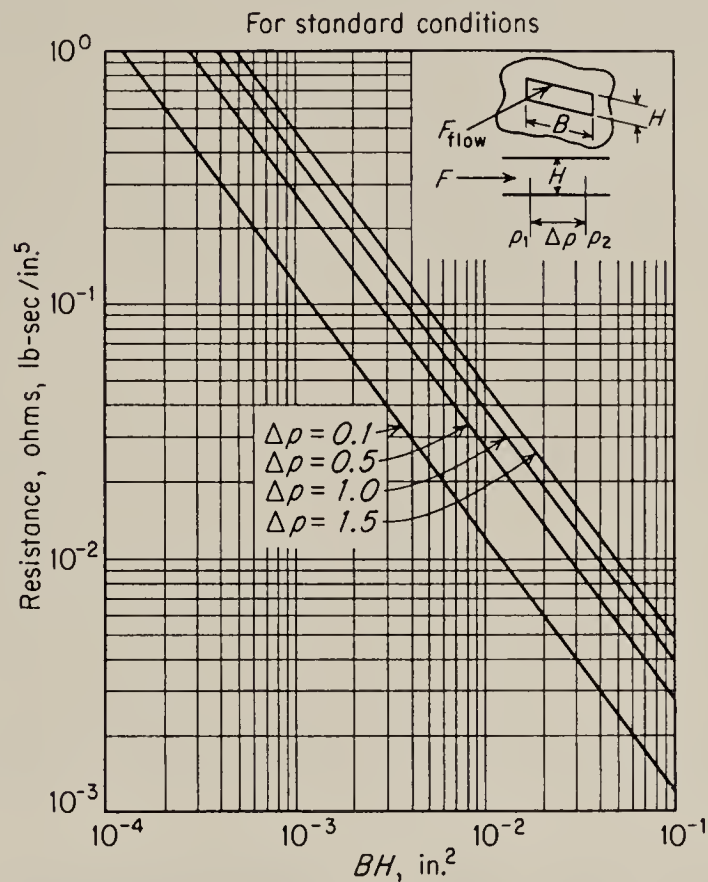


FIG. 12.9. Resistance of rectangular orifice. (Stolarik)

and pressure. For both Fig. 12.8 and Fig. 12.9 a correction factor for other than standard conditions may be derived as

$$k_{pT} = \left(\frac{P}{14.7}\right)^{\frac{1}{2}} \left(\frac{288}{T}\right)^{\frac{1}{2}} \tag{12.25}$$

For Figs. 12.10 and 12.11 the correction factor

$$k_T = \left(\frac{T}{288}\right)^{\frac{3}{4}} \tag{12.26}$$

may be applied for other than standard conditions.

The pneumatic capacitance of a volume can likewise be defined by analogy to electric circuits as

$$C = \int \frac{F \, dt}{\Delta P} \quad \text{in.⁵/lb} \tag{12.27}$$

where *F* is the volumetric flow in cubic inches per second and *P* is the pressure in pounds per square inch. The value of capacitance is independent of temperature but depends upon pressure and on the process of expansion. If the expansion is rapid, the process may be assumed to be

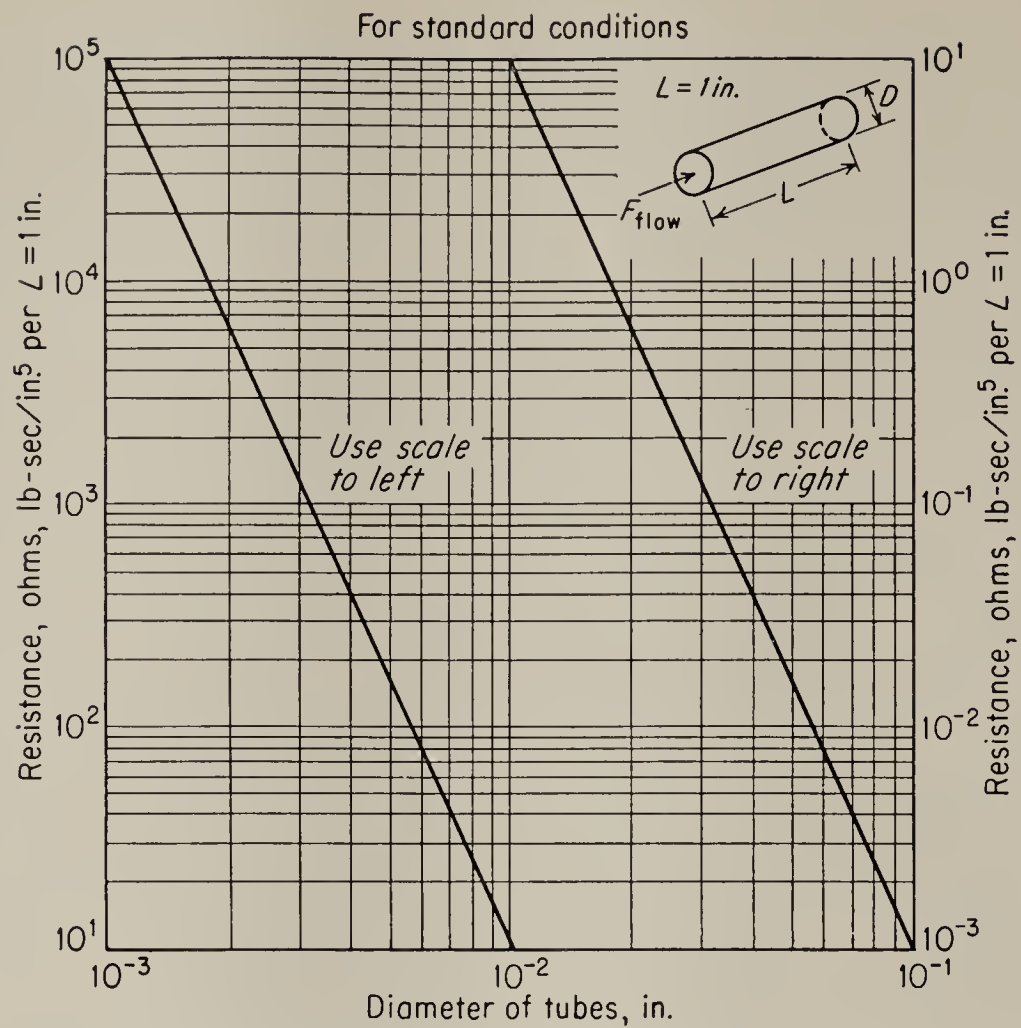


FIG. 12.10. Resistance of capillary tube. (Stolarik)

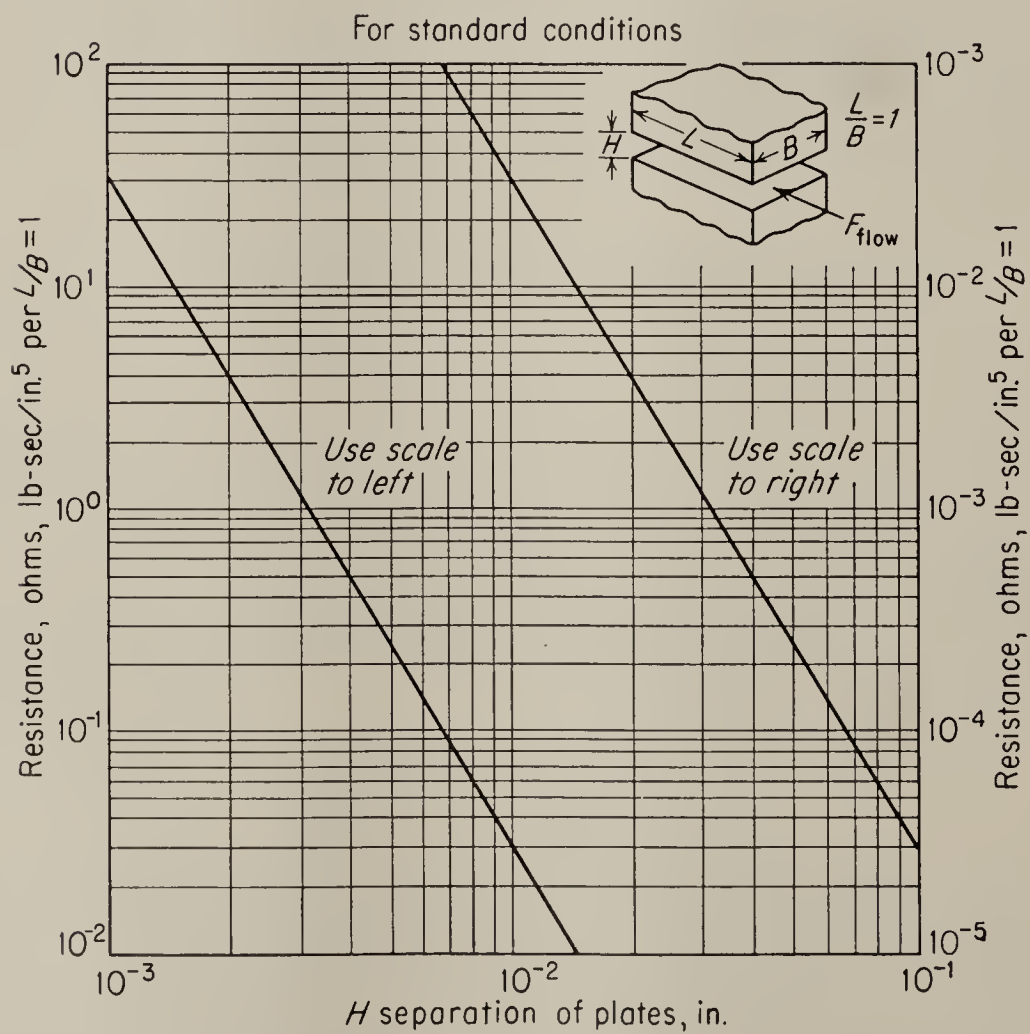


FIG. 12.11. Resistance of flat plates. (Stolarik)

adiabatic, and if the expansion is very slow, an isothermal process may be assumed. Usually the actual process is somewhere between the two limits. Figure 12.12 relates capacitance to volume, and Fig. 12.13 gives the correction for pressure.

In addition to their use in the synthesis of pneumatic control functions, the relations between physical configuration and pneumatic resistance and capacitance permit the determination of time constants of a pneumatic element that performs an operational function in the loop, such as the bellows in the example in the following section.

A second method of synthesizing control functions is to design a pneumatic or pneumatic-mechanical closed-loop system with the required

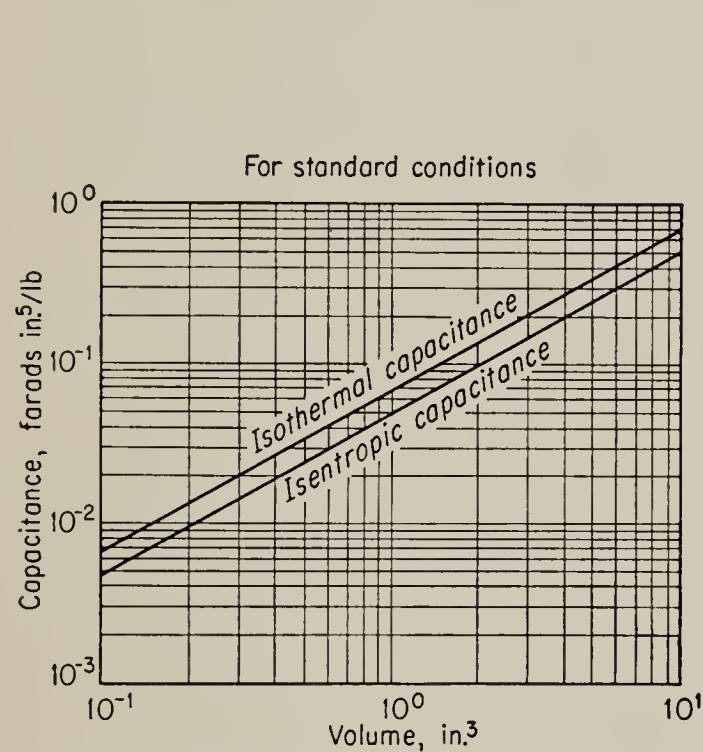


FIG. 12.12. Capacitance. (Stolarik)

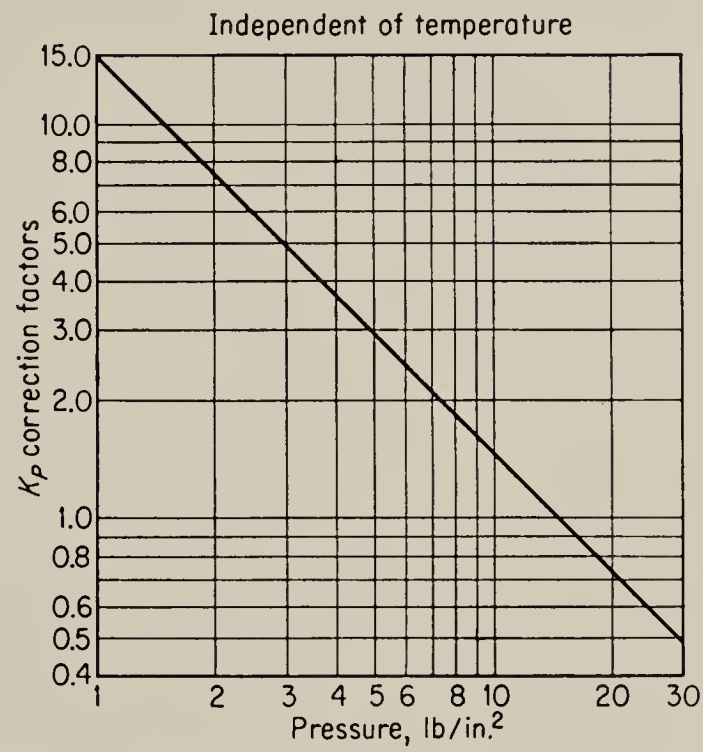


FIG. 12.13. Pressure correction for capacitance. (Stolarik)

input-output relations. Caldwell¹ has compiled a table of such units that is reproduced in Fig. 12.14. Williamson² has designed small, compact, adjustable lag networks and lead networks in the modular, or “plug-in,” form for the same types of control applications. Figure 12.15 shows sketches of these devices.

As an example of the analysis of a pneumatic equalizer, let us consider the lag-lead network shown in Fig. 12.16.* We desire the transfer function from flapper input position θ to output pressure p . We shall assume

¹ W. I. Caldwell, Generating Control Functions Pneumatically, *Control Eng.*, vol. 1, no. 1, p. 58, 1954.

² H. Williamson, Theory and Design of Compound Action Pneumatic Controllers, *Trans. Soc. Instrument Tech.*, vol. 6, no. 4, p. 153, 1954.

* A. R. Aikman and C. I. Rutherford, The Characteristics of Air Operated Controllers, in Tustin (ed.), “Automatic and Manual Control,” Proceedings of Cranfield Conference, 1951, Butterworth & Co. (Publishers) Ltd., London, 1952, p. 175.

PROPORTIONAL PLUS RATE

Basic Circuit	Operational Equation	Magnitude Ratio	Remarks
	<p>where</p> $\frac{P - P_0}{\theta - \theta_s} = \frac{A_1}{A_2} \left[\frac{1 + Tp}{1 + (\epsilon + A_3/A_2)Tp} \right]$ <p>if</p> $\frac{1}{\epsilon + A_3/A_2} = \frac{A_2}{A_3} = a$ <p>$\epsilon \ll A_2/A_3$</p>		<p>At low frequencies P_f follows P and M is equal to A_1/A_2. At high frequencies little air bleeds through R and M equals $aA_1/A_2 = (A_1/A_2)^3$. The controller response depends on the fixed design constant A. Maximum lead time of ramp response, maximum phase lead, and maximum increase in ratio of amplitude varies with a. When a equals unity, response corresponds to that of a proportional-position circuit.</p>
	<p>if</p> $\frac{P - P_0}{\theta - \theta_s} = \frac{A_1}{A_2} \left[\frac{1 + Tp}{1 + Tp/a} \right]$ <p>$\epsilon \ll 1$ and $\epsilon \ll \frac{1}{a}$</p> <p>where $1/a$ is the change in P_f for a unit change in P when R is completely closed</p>		<p>At low frequencies P_f follows P, and M equals A_1/A_2. At high frequencies little air bleeds through R. However, elongation of bellows with increase in P causes a corresponding but larger change in P_f. This provides direct feedback which limits the value of M to aA_1/A_2.</p>
	<p>where</p> $\frac{P - P_0}{\theta - \theta_s} = \frac{A_1}{A_2} \left[\frac{1 + (A_3/A_2)T_1p}{1 + T_1p} \right]$ <p>if</p> $T = \frac{A_3}{A_1} T_1 \quad s = A_3 A_1$ <p>$\epsilon \ll 1$</p>		<p>At low frequencies θ_f essentially follows θ and the value of M is A_1/A_2. At high frequencies little air bleeds through R, θ_f remains essentially constant and M equals $aA_1/A_2 = (A_3/A_2)$.</p>

PROPORTIONAL PLUS RESET

	<p>where</p> $\frac{P - P_0}{\theta - \theta_s} = \frac{A_1}{A_2} \left[\frac{1 + 1/Tp}{1 + (1 + \epsilon - A_3/A_2)/Tp} \right]$ <p>if</p> $U = 1/T$ <p>$b = 1 + \epsilon - A_3/A_2$</p> <p>(By adjustment of A_3/A_2, b can be made small and b/Tp term neglected)</p>		<p>At low frequencies P_f about equals P and the negative feedback essentially cancels the positive feedback. Consequently small changes in $\theta - \theta_s$ cause large changes in P. At high frequencies, P_f is essentially constant so that M is equal to A_1/A_2.</p>
--	---	--	--

Basic Circuit	Operational Equation	Magnitude Ratio	Remarks
	$\frac{P - P_0}{\theta - \theta_s} = \frac{A_1}{A_2 k} \left[\frac{1 + 1/Tp}{1 + \epsilon/kTp} \right], \text{ if } \epsilon \ll k$ $U = 1/T$ $b = \epsilon/k$ $k = \text{change in } P_f \text{ for a unit change in } P \text{ when } R \text{ is closed}$		<p>At low frequencies P_f is almost equal to atmospheric pressure, there is no feedback, and the value of M is high. At high frequencies the action of P on the bellows causes a corresponding change in P_f. The resultant feedback reduces the amplitude of P.</p>
	$\frac{P - P_0}{\theta - \theta_s} = \frac{A_1}{A_2} \left(1 + \frac{R_1}{R_2} \right) \left(\frac{1}{Tp} + 1 \right) \left(1 + \frac{R_1 Cp}{1 + R_1/R_2} \right)$ <p>where</p> $\epsilon' = 1 + \epsilon - A_3/A_2$ <p>When operating frequency is between $\frac{1}{2\pi T}$ and $\frac{1}{2\pi R_1 C}$</p> <p>then</p> $U = 1/T$ $B = \frac{A_2}{A_1} \frac{R_2}{R_1 + R_2} \frac{\Delta P}{\Delta \theta}$		<p>The pressure divider $R_1 R_2$ proportions the positive feedback pressure between the output pressure P and the lagged pressure P_m. The amount of cancellation of the negative feedback by the action of P on the negative feedback pressure and the proportional response of the controller depends on the ratio R_1/R_2. The one-to-one relay applies the pressure P_m in the capacity tank to the pressure divider circuit without permitting flow between them. This makes the reset rate, $1/T$, independent of the values of R_1 and R_2. By modifying the circuit so that a reducing valve supplies P_m to the pressure divider, the circuit has proportional-position action with the bandwidth adjustable by changing the ratio R_1/R_2.</p>
	$\frac{P - P_0}{\theta - \theta_s} = \frac{A_1}{A_2} \left(1 + \frac{R_1}{R_2} \right) \left[\frac{1}{(R_1 + R_2)C_1 p} + 1 + \frac{R_1 R_2 C_2 p}{(R_1 + R_2)} \right]$ <p>if $C_2 \ll C_1$ and $\epsilon \ll R_2/R_1$</p> <p>When the operating frequency is between $\frac{1}{2\pi(R_1 + R_2)C_1}$ and $\frac{1}{2\pi R_1 R_2}$</p> <p>then $U = \frac{1}{(R_1 + R_2)C_1}$ and</p> $B = \frac{\Delta P}{\Delta \theta} \frac{A_2}{A_1} \frac{R_2}{R_1 + R_2}$		<p>The pressure divider circuit controls the proportional bandwidth and the reset rate. When the value of either resistance is changed, the values of both the bandwidth and reset rate are changed.</p>

FIG. 12.14. Pneumatic-mechanical equalizers.

▲ indicates supply pressure.

PROPORTIONAL PLUS RESET PLUS RATE

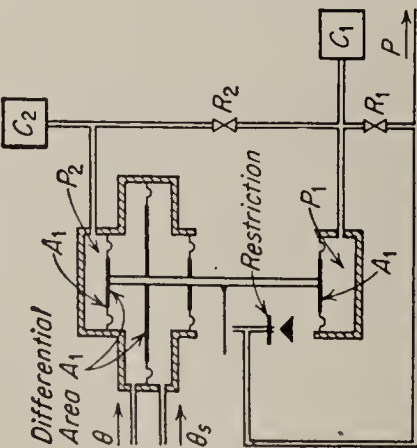
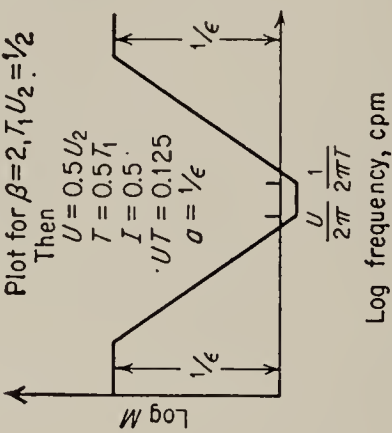
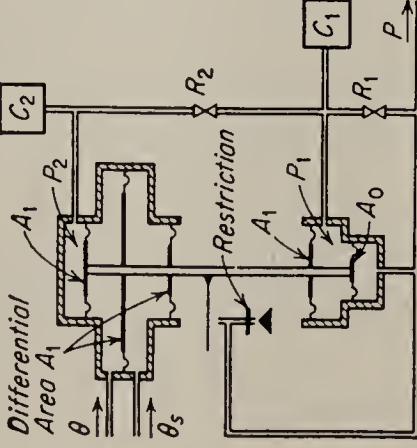
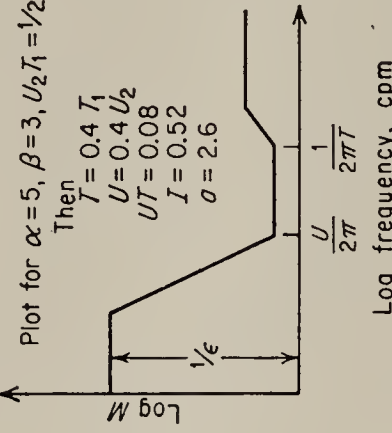
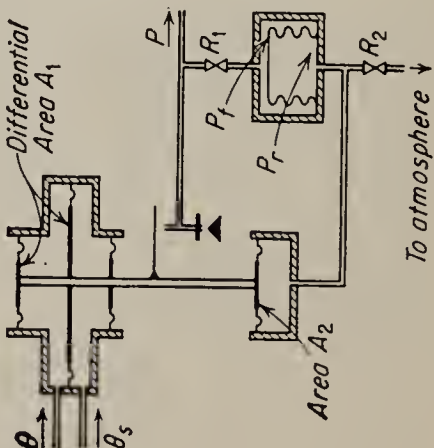
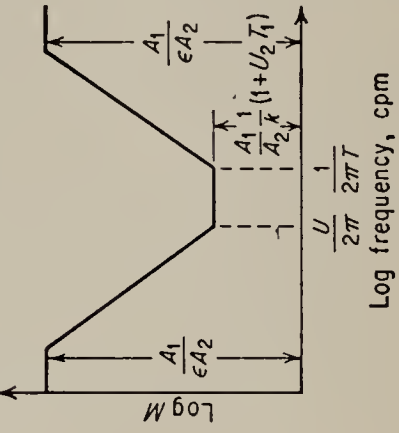
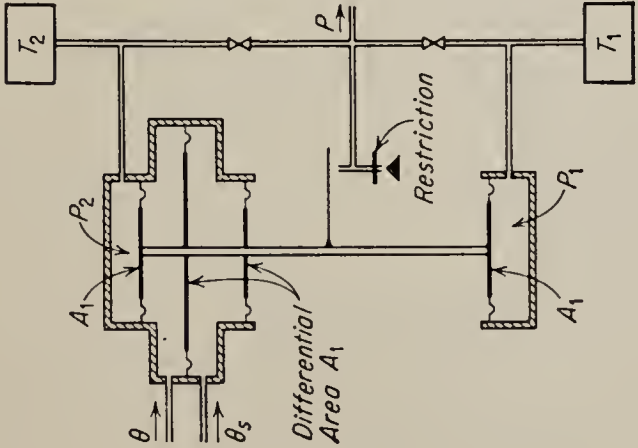
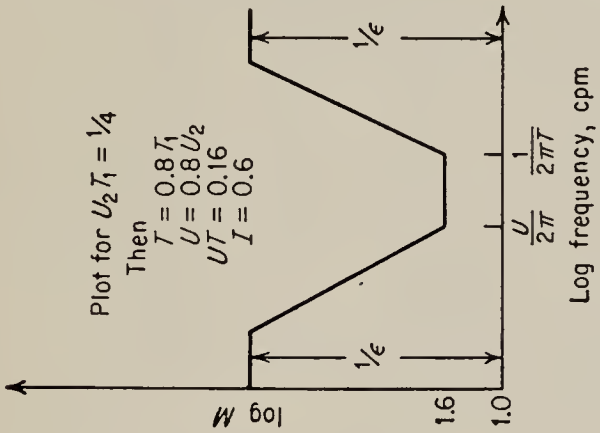
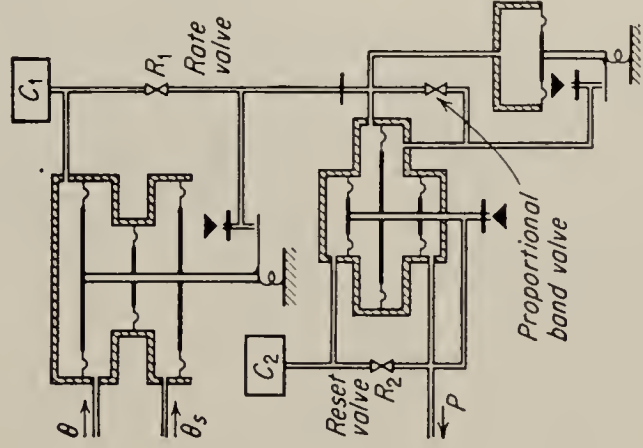
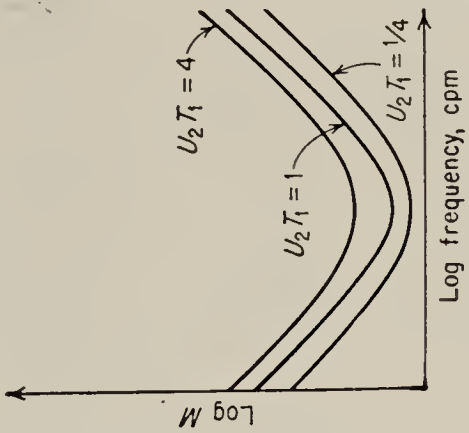
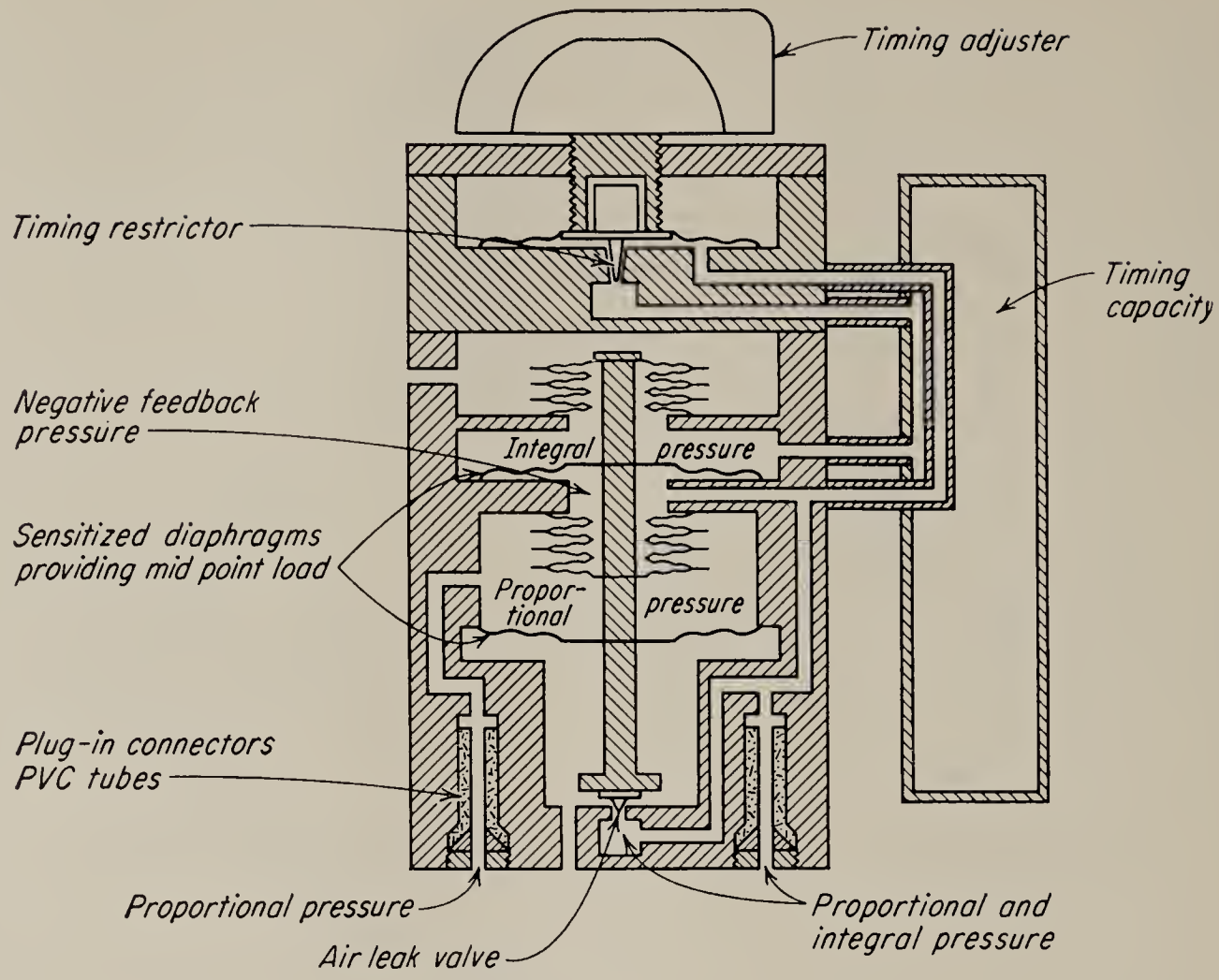
Basic Circuit	Operational Equation	Magnitude Ratio	Remarks
	$\frac{P - P_0}{\theta - \theta_s} = (1 + \beta T_1 U_2) \left[\frac{\frac{U_2/p}{1 + \beta T_1 U_2} + 1 + \frac{T_1 p}{1 + \beta T_1 U_2}}{\epsilon U_2/p + 1 + \epsilon T_1 p} \right]$ <p>where $\epsilon \beta T_1 U_2 \ll 1$ $T_1 = R_1 C_1$ $U_2 = 1/R_2 C_2$ $\beta = 1 + C_2/C_1$</p>	 <p>Plot for $\beta=2, T_1 U_2=1/2$ Then $U=0.5 U_2$ $T=0.5 T_1$ $I=0.5$ $UT=0.125$ $a=1/\epsilon$</p>	
	$\frac{P - P_0}{\theta - \theta_s} = \frac{1 + U_2 T_1 \beta}{1 + U_2 T_1 \beta/a} \left[\frac{\frac{U_2/p}{1 + U_2 T_1} + 1 + \frac{T_1 p}{1 + U_2 T_1 \beta}}{\epsilon U_2/p + 1 + \frac{(T_1/a)p}{1 + U_2 T_1 \beta/a}} \right]$ <p>where $a = A_1 A_0$ $\beta = 1 + C_2/C_1$ $\ll 1$ Then $T = T_1/(1 + U_2 T_1)$ $U = U_2/(1 + U_2 T_1)$</p>	 <p>Plot for $\alpha=5, \beta=3, U_2 T_1=1/2$ Then $T=0.4 T_1$ $U=0.4 U_2$ $UT=0.08$ $I=0.52$ $a=2.6$</p>	
	$\frac{P - P_0}{\theta - \theta_r} = \frac{A_1}{A_2 k} [1 + U_2 T_1] \left[\frac{\frac{U_2/p}{1 + U_2 T_1} + 1 + \frac{T_1 p}{U_2/kp + 1 + (T_1/k)p}}{U_2/kp + 1 + (T_1/k)p} \right]$ <p>assuming $k \ll 1$ and $\epsilon \ll k/(1 + U_2 T_1)$ k is the ratio of change of P_r to P_f when $R_2 = \infty$ $U = U_2/(1 + U_2 T_1)$ $T = T_1/(1 + U_2 T_1)$ $I = k/(1 + U_2 T_1)$</p>	 <p>Log frequency, cpm</p>	

FIG. 12.14. (Continued)

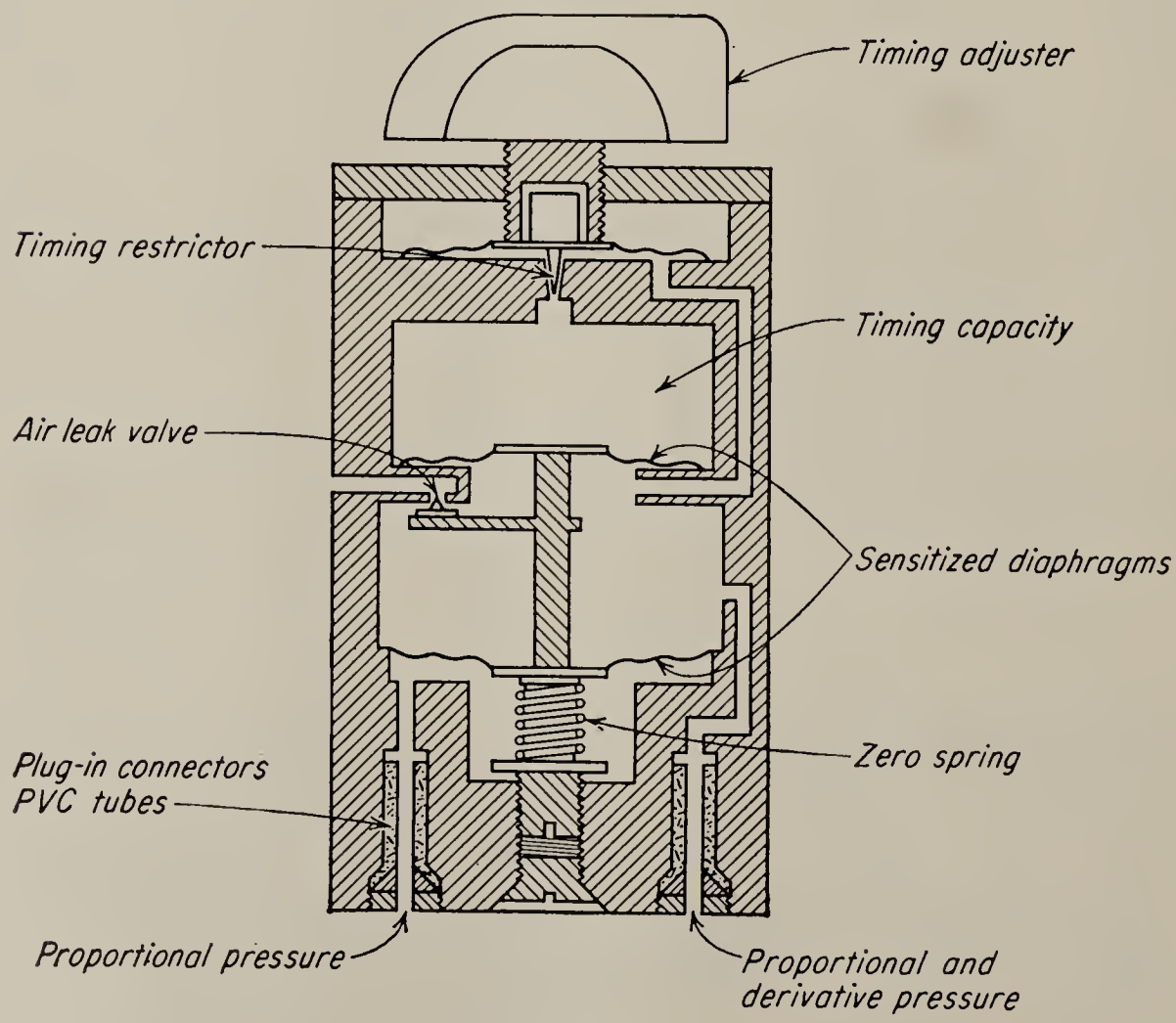
PROPORTIONAL PLUS RESET PLUS RATE

Basic Circuit	Operational Equation	Magnitude Ratio	Remarks
	$\frac{P - P_0}{\theta - \theta_s} = \frac{1 + U_2 T_1}{1 - U_2 T_1} \left[\frac{U_2/p}{1 + U_2 T_1} + 1 + \frac{T_1 p}{1 + U_2 T_1} \right]$ <p>where $\epsilon \ll 1$ $U_2 = 1/T_2$ Then T_1 = rate time when $U_2 = 0$ ($T_2 = \infty$) U_2 = reset rate when $T_1 = 0$ Using both reset and rate $T = T_1 (1 + U_2 T_1)$ = rate time $U = U_2 / (1 + U_2 T_1)$ = reset rate and the proportional band is decreased by</p> $I = \frac{1 - U_2 T_1}{1 + U_2 T_1}$ <p>where I = interaction factor</p>		<p>With this circuit the largest practical value of $U_2 T_1$ is about 0.25.</p>
	$\frac{P - P_0}{\theta - \theta_s} = S_1 (1 + U_2 T_1) \left[\frac{U_2/p}{1 + U_2 T_1} + 1 + \frac{T_1 p}{1 + U_2 T_1} \right]$ <p>where S_1 = proportional amplification term dependent on the setting of the valve in the amplification stage $T_1 = R_1 C_1$ $U_2 = \frac{1}{R_2 C_2}$ Effective values are: $U = \frac{U_2}{1 + U_2 T_1} = \frac{1/T_1}{1 + 1/U_2 T_1}$ $T = \frac{T_1}{1 + U_2 T_1} = \frac{1/U_2}{1 + 1/U_2 T_1}$ $I = 1/(1 + U_2 T_1)$ When $U_2 T_1$ is small, $U \cong U_2$, $T \cong T_1$ When $U_2 T_1 \gg 1$, $U \cong \frac{1}{T_1}$, $T \cong \frac{1}{U_2}$</p>		<p>By proper adjustment, input to reset stage can be made to lead controlled variable by sufficient time to bring process under control without overshoot. Response curves are normal for normal load changes.</p>

I = interaction, a term which indicates how much the proportional band of a controller is increased by the interaction of the reset and rate effect. ▲ indicates supply pressure.



(a) Lag



(b) Lead

FIG. 12.15: Plug-in pneumatic equalizers. (Williamson)

that the flapper is operating in its linear range. Thus

$$p = P - \frac{b}{a+b} c\theta - \frac{a}{a+b} cKp_2 \quad (12.28)$$

where P = initial output pressure

c = gain constant of flapper, psi/in.

K = spring constant of feedback bellows, in./psi

Following the usual procedure of writing the linear equation about an operating point, P may be struck out of Eq. (12.28). For the derivative restrictor we may write

$$\hat{p} = (1 + T_1 s) \hat{p}_1 \quad (12.29)$$

where the constant T has the dimensions of a time constant. Its value is determined by the area of the restriction and the volume of air at the

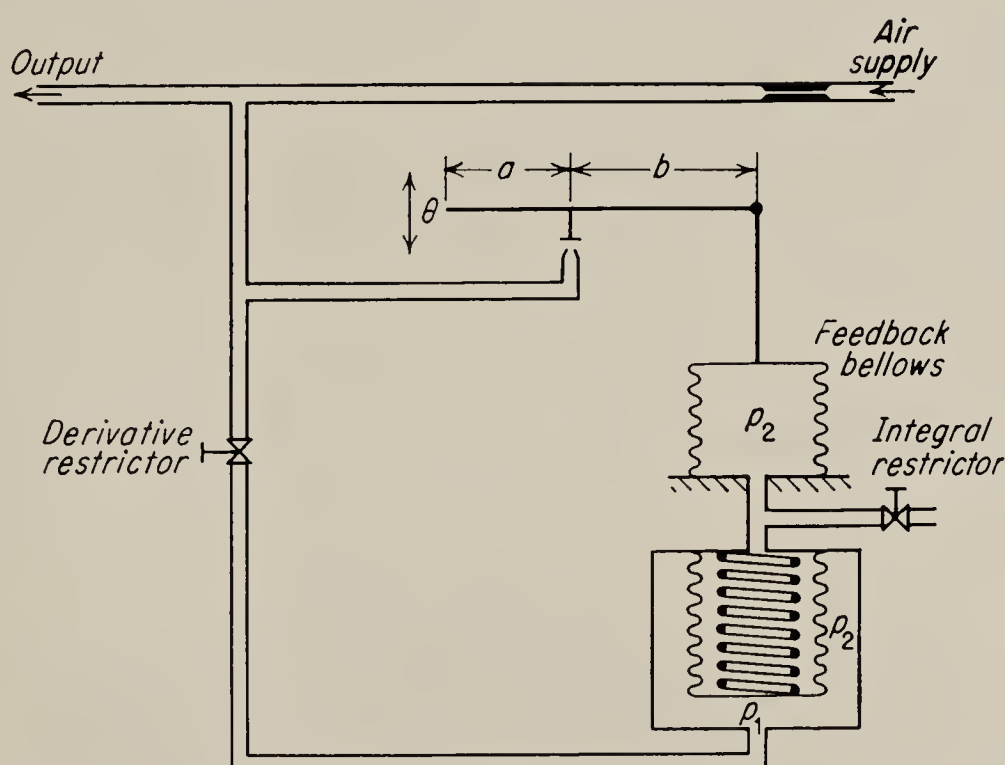


FIG. 12.16. A pneumatic lag-lead network, or a so-called type I controller.

pressure p_1 . Likewise we may relate p_1 and p_2 through the second restrictor by

$$AT_2 s \hat{p}_1 = (1 + T_2 s) \hat{p}_2 \quad (12.30)$$

where $A = \frac{\partial p_2}{\partial p_1}$ at operating point

T_2 = time constant of integral constriction and volume of feedback bellows

Solving the three equations, we obtain

$$\frac{\hat{p}}{\hat{\theta}} = k_2 \frac{(T_1 s + 1)(T_2 s + 1)}{T_1 T_2 s^2 + (T_1 + T_2 + k_1 T_2) s + 1} \quad (12.31)$$

where $k_1 = \frac{acKA}{a+b}$ and $k_2 = -\frac{bc}{a+b}$

The form of Eq. (12.31) is rather different from that commonly used in the process-control industries. There, the traditional method of analysis would relate p and θ by the differential equation containing the sum of the derivative, proportional, and integral terms. An attempt is usually made to cling to a physical interpretation of the separate terms by the introduction of interaction terms, which turn out to be various ratios of the time constants in Eq. (12.31). The interaction terms attempt to relate controller dial settings, which adjust the two restrictors in the example above, to the time constants¹ of the transfer function. As the advantages of operational techniques become more widely known, the transfer function and the frequency response of elements will themselves become a part of physical reality for process-control engineers, and manipulation of the time constants will be acceptable. This linearized analysis may be used to develop transfer functions for the configurations shown in Fig. 12.15 and the additional configurations given in the problems at the end of this chapter.

12.8. A Pneumatic Control System. As an example of a pneumatic control system that includes several of the more common types of components, we shall consider the pneumatic positioning system shown in Fig.

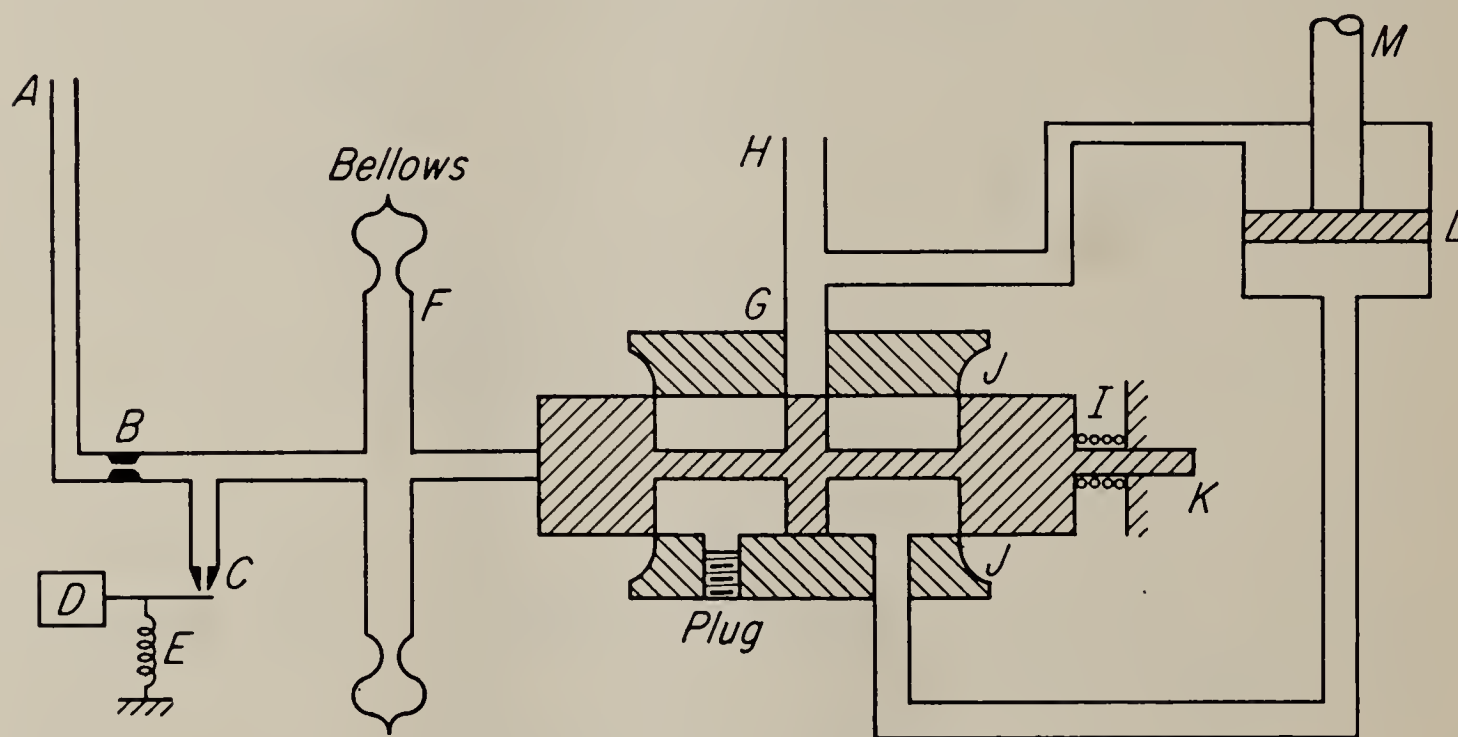


FIG. 12.17. A pneumatic positioning system.

12.17. This system might be used to actuate the control surfaces of an aircraft or to perform any other positioning operation. A is a source of low-pressure air (50 psi) for operating the flapper valve. This low-pressure supply, in contrast to the high-pressure source, must have its pressure rather closely regulated, since the spring E is adjusted for a given pressure. Likewise the spring I is adjusted so that the entire system will be in the center of its operating range with no signal to the torque motor.

¹ J. Janssen, *Analysis of Pneumatic Controllers*, in Tustin (ed.), *ibid.*, p. 189.

It would be possible to operate the flapper valve on high-pressure air, but the leakage and the force required from the torque motor would be unnecessarily high, and furthermore it is difficult to regulate the pressure properly. B is the upstream orifice, and C is the downstream or controlled orifice of the flapper valve. D is the torque motor that actuates the flapper. The spring E balances the force from the air jet. The bellows F operates the pilot valve G against the spring I . H is the high-pressure air supply (500 to 1,000 psi). The exhaust port J of the pilot valve is larger than the inlet port, to allow the expanded low-pressure air to escape without throttling. This is one of the few particulars in which pneumatic valves differ from hydraulic valves. The extension of the pilot-valve spool K may be used if it is desirable to pick off spool position. This two-stage pneumatic amplifier consisting of the flapper valve and pilot valve would be referred to as a "relay" in the process-control industry. It drives the power piston L , which operates on the differential-area principle. Even if the pressure were the same on both sides of the power piston, it would move upwards, because the pressure acts over a larger area on one side of the piston. The output position M actuates the load and is normally fed back to close the control loop.

From the data in Chap. 13 we shall see that in the operating range there is a linear relationship between flapper position and bellows pressure. The bellows force tending to open the valve will be assumed to be proportional to the chamber gauge pressure. Newton's law for the mechanical system is thus

$$\hat{\mathcal{F}}_{\text{bellows}} = (Ms^2 + Bs + k_1 + k_b + k_c)x \quad (12.32)$$

where M is the mass of the moving parts and B is the coefficient of viscous damping. k_1 is the coefficient of the spring I ; k_b is the spring constant of the bellows; and k_c is the equivalent spring due to the Bernoulli force at the valve orifice. Thus the transfer function from bellows force to valve position is

$$\frac{\hat{x}}{\mathcal{F}_{\text{bellows}}} = \frac{1}{Ms^2 + Bs + k_1 + k_b + k_c} \quad (12.33)$$

The relationship between flapper position and bellows force may be assumed linear in the operating range; however there is a pneumatic time lag between flapper position and bellows pressure which is proportional to bellows force

$$\frac{\hat{\mathcal{F}}_{\text{bellows}}}{\hat{x}_{\text{flap}}} = \frac{k_{\text{flap}}}{1 + T_b s} \quad (12.34)$$

The method of finding T_b for a given system is discussed below. The transfer function from valve opening to pressure on the power piston depends on the length of the transmission line. If the line is short, the

transmission-line effects may be ignored, and a single time constant made up of the resistance of the valve orifice and the capacitance due to the volume of air in the line and the power piston may be assumed. If the line is not short, the approximation of a transport lag and the time constant, as discussed in Sec. 12.6, will usually be adequate.

A simple equivalent circuit may be employed for the pneumatic valve if the valve may be assumed to be approximately linear. This circuit is

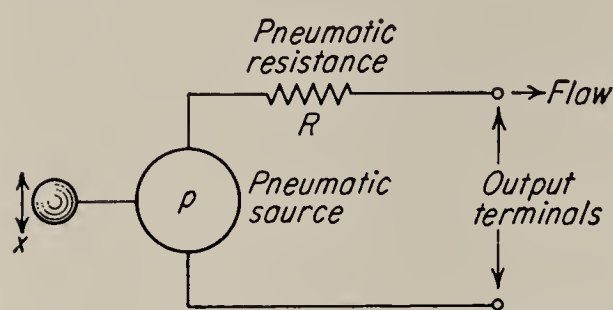


FIG. 12.18. Equivalent circuit of pneumatic valve.

shown in Fig. 12.18. According to this circuit the pressure p delivered by the valve for zero (or constant) flow is proportional to x , the valve spool position. The slope $\partial p / \partial x$ is referred to as the pressure sensitivity of the valve, and may be obtained by direct measurement, or by calculations of the sort performed in Chap. 11 for hydraulic systems. As a

result of fluid flow to the load this pressure is reduced, and this effect is indicated by the series resistor R . This resistance, which is the slope of the pressure versus flow curve, may also be obtained by measurement or by calculation. The reader may recognize this equivalent circuit of the valve as being in the form of the Thévenin equivalent¹ circuit used in electric circuit analysis.

The time constant T_b will now be determined, although it will be found to be a function of the operating conditions rather than a true constant.²

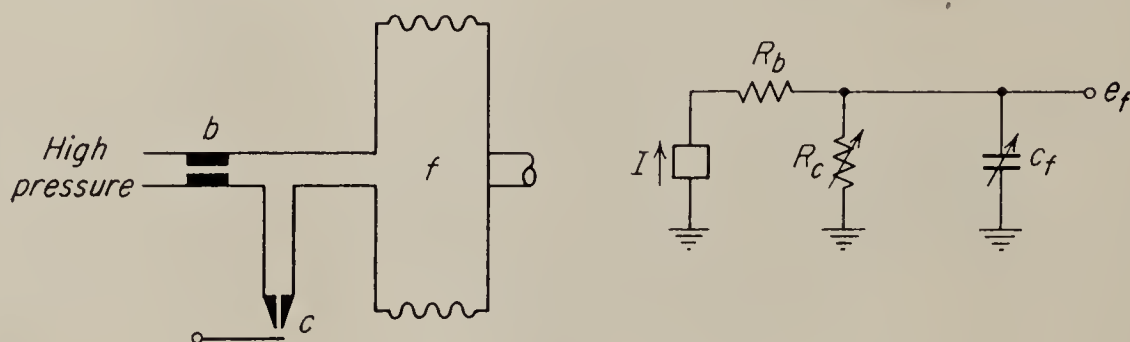


FIG. 12.19. The flapper valve and its electric analogue.

Figure 12.19 shows the pilot valve and its electric analogue. In the normal operating range the ratio of the upstream pressure to the downstream pressure across orifice B is greater than 2:1; thus the flow through the orifice is constant. In the electric analogue this is represented by a constant current source I . The transfer function of interest is from a change in the flapper orifice resistance R_c to the change in bellows pressure, repre-

¹ W. Everitt and G. Anner, "Communication Engineering," 3d. ed., McGraw-Hill Book Company, Inc., New York, 1956, p. 231.

² For a derivation that assumes linearity about an operating point, see H. A. Helm, Frequency Response Approach to the Design of a Mechanical Servo, *Trans. ASME*, vol. 76, p. 1209, 1954, and the appendix by R. Oldenburger.

sented by the voltage e_F . Unfortunately, as the bellows expands, its equivalent capacitance changes; thus C_F is a function of x . If the change in bellows volume is only a small percentage of the total bellows volume, this variation in C may be neglected. We may write by Kirchhoff's law that

$$I = i_{R_C} + i_{C_F} \quad (12.35)$$

$$\text{or} \quad I = \frac{e_F}{R_C} + C_F \frac{de_F}{dt} \quad (12.36)$$

Taking the Laplace transform we obtain

$$\frac{I}{s} = \frac{\hat{e}_F}{R_C} + C_F s \hat{e}_F - C_F e_F(0) \quad (12.37)$$

where $e_F(0)$ is the initial value of the voltage e_F . Suppose now that at time $t = 0$ the resistance R_C is changed suddenly from the value R_{C1} to a new value R_{C2} . If we assume that a static condition exists prior to this change then de_F/dt in Eq. (12.36) is zero, and the initial value of the voltage on the capacitance is

$$e_F(0) = R_{C1}I \quad (12.38)$$

Substituting this value for $e_F(0)$ in Eq. (12.37) and letting R_C take on the new value R_{C2} we obtain

$$\frac{I}{s} = \frac{\hat{e}_F}{R_{C2}} + C_F s \hat{e}_F - R_{C1}C_F I \quad (12.39)$$

Solving for \hat{e}_F gives

$$\hat{e}_F = \frac{R_{C2}I(T_1s + 1)}{s(T_2s + 1)} \quad (12.40)$$

where

$$\begin{aligned} T_1 &= R_{C1}C_F \\ T_2 &= R_{C2}C_F \end{aligned} \quad (12.41)$$

Finally if we take the inverse Laplace transform of Eq. (12.40) we find that

$$e_F(t) = IR_{C2} + I(R_{C1} - R_{C2})e^{-t/T_2} \quad (12.42)$$

Thus the voltage e_F approaches its final value in the familiar exponential manner, and the time constant of the circuit is T_2 . When R_c is varied in some other manner than stepwise, T_2 varies during the process; however, as an approximation for the time constant, some average value of resistance may be assumed.

PROBLEMS

12.1. Find the mass weight of flow for a smoothly rounded orifice $\frac{1}{16}$ in. in diameter at 20°C. The upstream pressure is 1,000 psi.

12.2. Find the equivalent time constant of a 5-in.-diameter cylinder 50 in. long with a 1-lb piston.

12.3. Find the physical dimensions of a pneumatic network at 15°C and 1,000 psi to synthesize the function $1/(Ts + 1)$.

12.4. In Fig. 12.20 is shown a schematic of a so-called type II controller, which is sometimes called a series controller because of the arrangement of the derivative and integral restrictors.¹ Find the transfer function from θ to p .

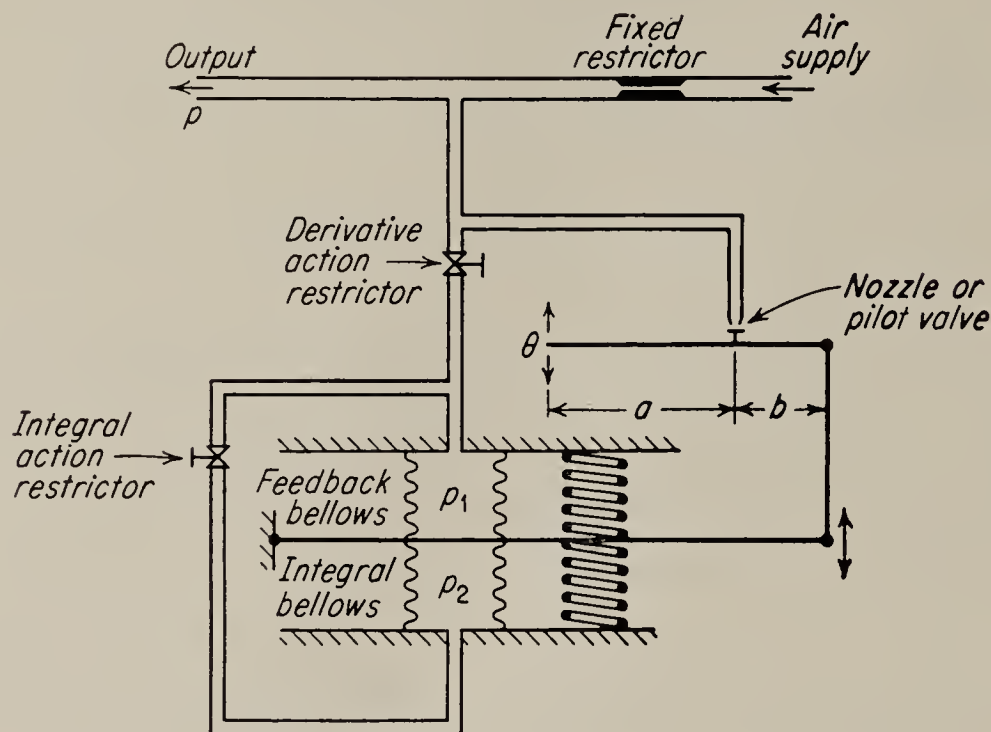


FIG. 12.20

12.5. Find the transfer function of the parallel type III controller shown in Fig. 12.21, which is taken from Aikman and Rutherford.

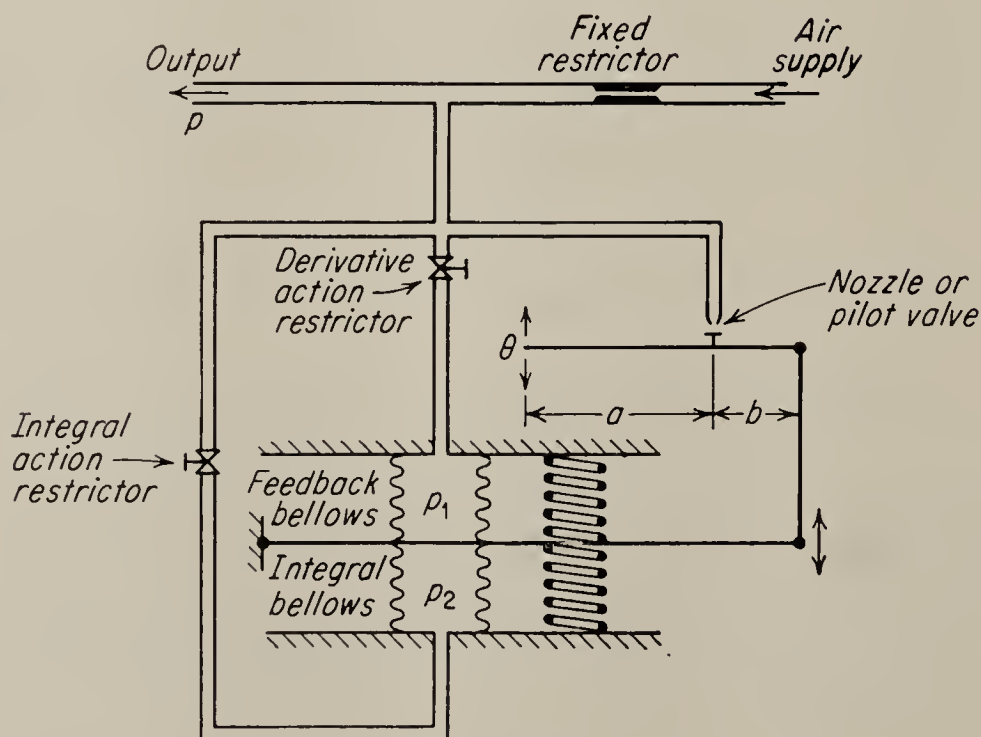


FIG. 12.21

¹ Aikman and Rutherford, *loc. cit.*

12.6. Find the transfer function of the type IV controller shown in Fig. 12.22, taken from Aikman and Rutherford. Note the use of a combination of hydraulic and pneumatic elements.

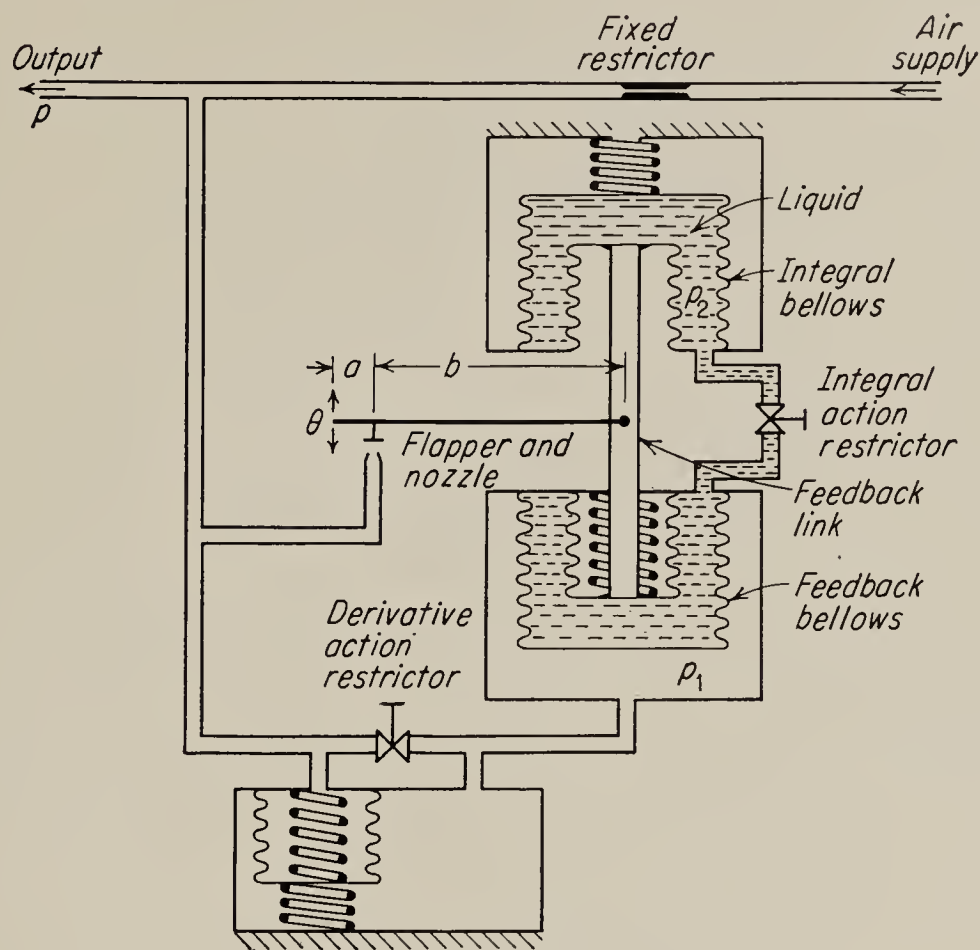


FIG. 12.22

12.7. Find the transfer function of the type V controller shown in Fig. 12.23, taken from Aikman and Rutherford.

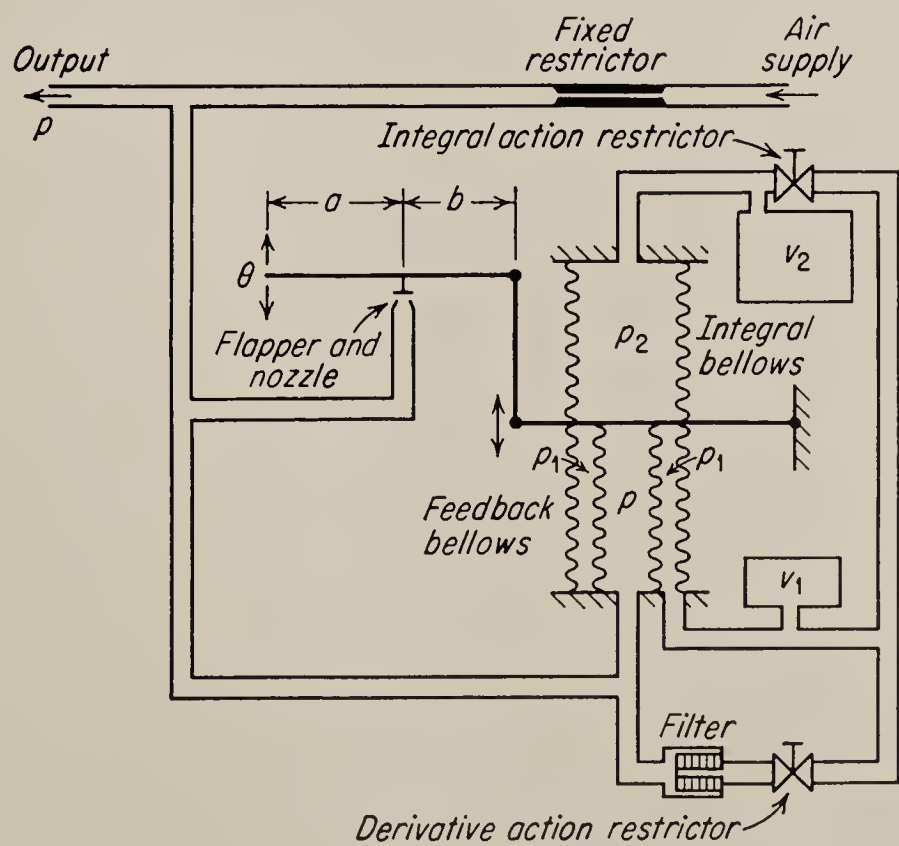


FIG. 12.23

12.8. Determine the transfer function of the lag-lead controller shown in Fig. 12.24, which is taken from Janssen.¹ Assume that the gain of the flapper-nozzle pilot valve is very high.

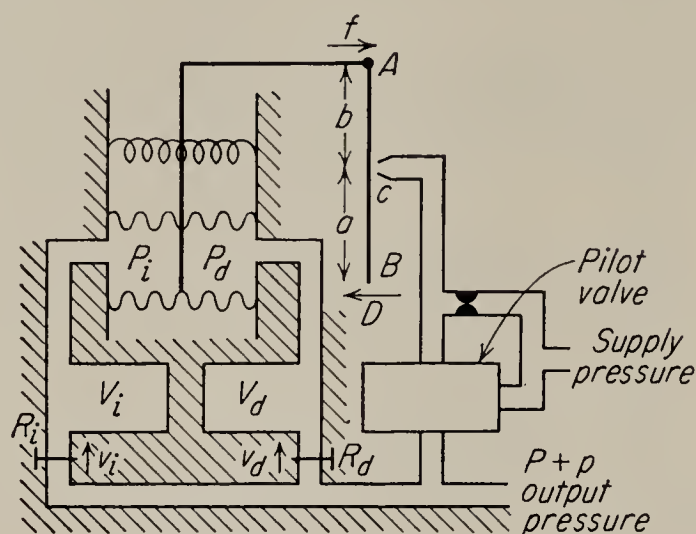


FIG. 12.24

12.9. Develop Eq. (12.16) from Eq. (12.15) by assuming constant weight rate of flow and constant temperature.

¹ Janssen, *op. cit.*

CHAPTER 13

PNEUMATIC COMPONENTS

13.1. Introduction. In general, all pneumatic control systems are of the valve-controlled type. There are few, if any, systems in which the pump is placed in the loop and directly controlled. Quite commonly the pneumatic pump is incapable of supplying the maximum load of the system by itself. Usually an accumulator, or storage element, and a pressure-regulating valve are included in the pressure-supply unit. In fact, in some units that have a limited operation time, e.g., in guided missiles, the pump is omitted altogether, and the system operates from an accumulator that is charged from a fixed supply before placing the system in operation.

Many of the components used in a high-pressure pneumatic system are identical with their hydraulic counterparts, and the two types may in fact be used interchangeably when this is true, as it is in several types of valves and motors, for example. One common difference in pneumatic components, however, is that the low-pressure outlets must usually be significantly larger than the inlet ports, so that the increased volume of gas at low pressure will not be throttled as it passes.

Pneumatic systems are usually slower than equivalent hydraulic systems owing to the finite time required for expansion of the gas. It is unfair, however, to compare the same system operating at the same supply pressure on the basis of speed of response with hydraulic and pneumatic fluids. The comparison should actually be made on a weight or cost basis for given specifications. Pneumatic systems are being designed to meet faster and faster response requirements, and certain pneumatic controls on jet aircraft, for example, have frequency responses that extend to higher than 10 radians/sec.

A typical example of an all-pneumatic control system on a military jet aircraft is shown in Fig. 13.1.¹ The device is employed to maintain a constant ratio² between the pressure in the combustion chamber and the

¹ W. E. Reed, Pneumatic Control of a Turbojet Variable Nozzle, *Control Eng.* vol. 3, p. 93, October, 1956.

² Actually there are reasons for adjusting this ratio as a function of flight condition in order to maintain constant turbine-blade temperature, etc.

downstream pressure in the turbojet engine. The differential pressure bellows controls the pneumatic valve, which in turn operates the actuator. Changing the nozzle opening changes p_2 , the downstream pressure. If the pressure p_2 is too large the valve stem is forced down and orifice 2 narrows. As a result p_3 increases, and equilibrium is reached when the force generated by the pressure difference $p_2 - p_3$ is equal to the opposing force produced by the bellows.

The bellows may be represented by a single lag with a time constant of about 0.01 sec in the example given, while the equivalent spring of the

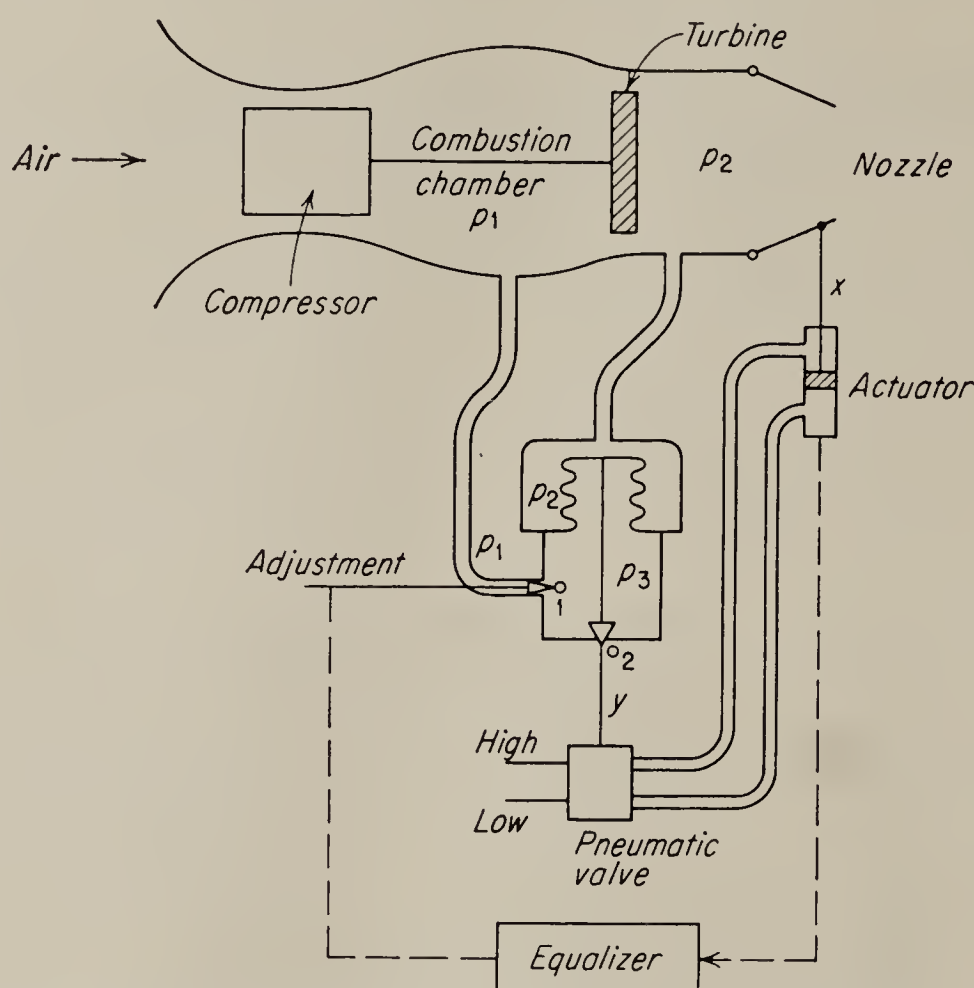


FIG. 13.1. An all-pneumatic servo in a jet-engine application. (Reed)

compressibility of the air in the actuator and the mass of the actuator plus the mass of the load contributes an underdamped complex-conjugate pair of roots. There is, in addition, the integration between valve position and actuator position. Reed finds the transfer from pressure ratio to actuator position from experimental data to be approximately

$$\frac{\hat{x}}{\Delta \hat{p}} = \frac{k}{s(0.01s + 1)(0.0004s^2 + 0.008s + 1)} \quad (13.1)$$

and a possible block diagram is shown in Fig. 13.2. It is seen from the equation that the system becomes unstable if k is too large. It is possible to equalize this system pneumatically by installing a compensator of the form shown in Fig. 13.3. The nozzle-position actuator is connected to the piston of the compensator, and the output of the compensator posi-

tions the orifice O_1 . The pressure in the volume C_1 is proportional to \dot{x} , and the force out is proportional to the rate of change of pressure on the top of the diaphragm. The actual linear transfer function for the equalizer is given in the figure. The usual linear design techniques may be used to set the time constants R_1C_1 and R_2C_2 so that the resultant frequency response of the system with equalizer indicates a satisfactorily stable system.

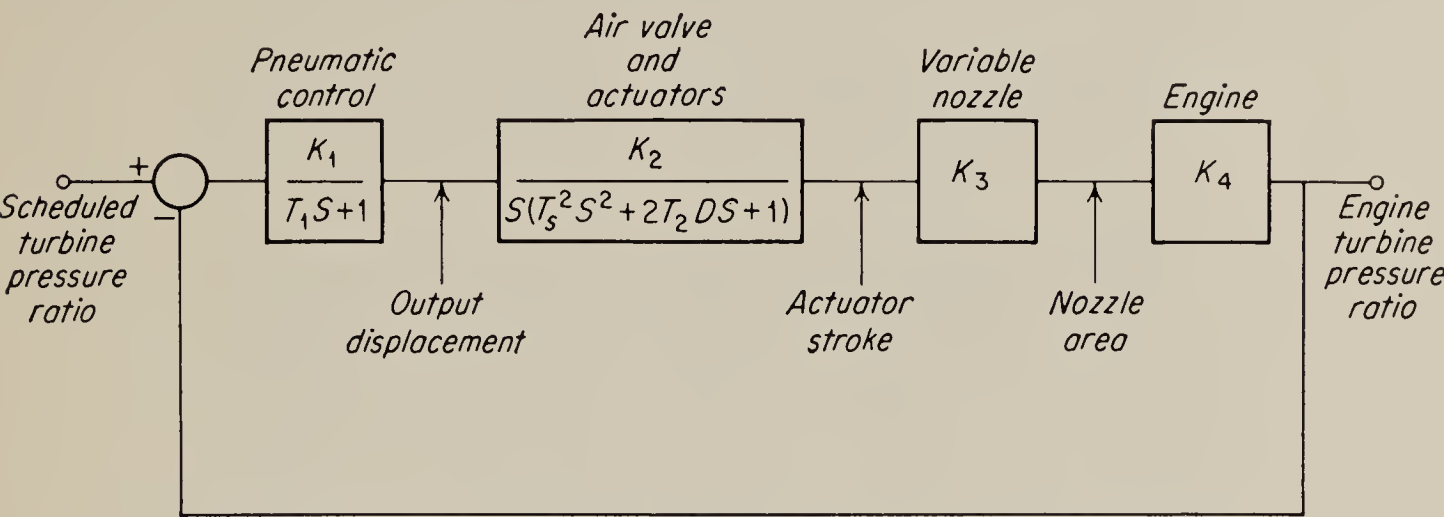


FIG. 13.2. Simplified block diagram of pneumatic control system.

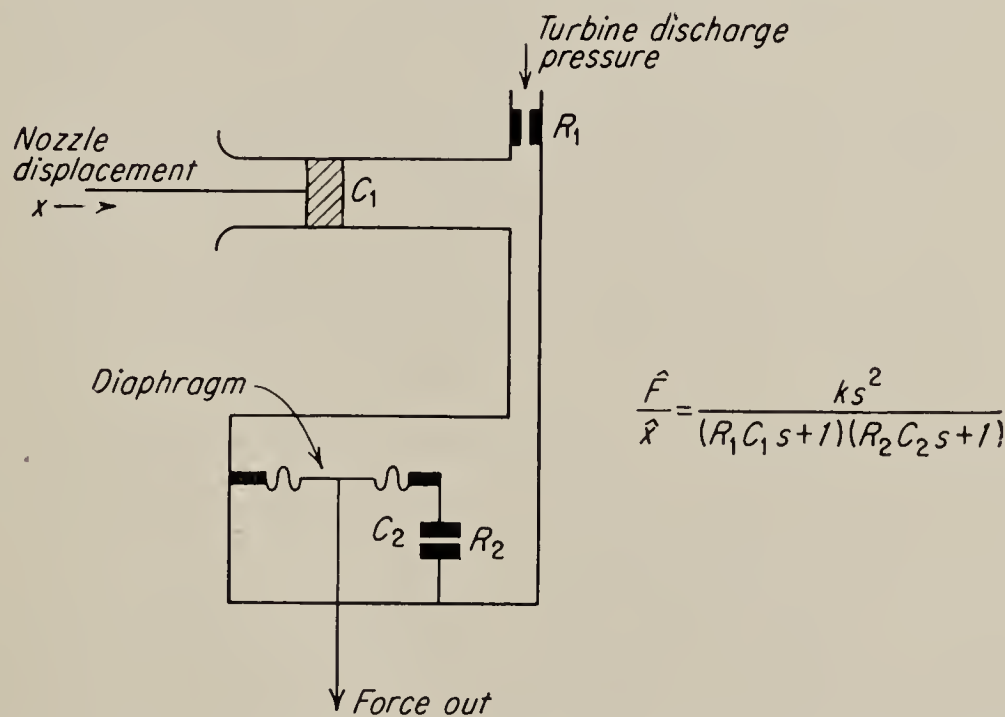


FIG. 13.3. A pneumatic equalizer for the system in Fig. 13.2.

The object of discussing this example is to point out that completely pneumatic control systems including equalizers have a definite place in modern high-speed applications. It is true, however, that at the present time the number of hydraulic installations far exceeds the number of pneumatic installations in high-speed, high-power applications.

13.2. Comparison of Weight of Electric, Hydraulic, and Pneumatic Systems. In Chap. 12 an aircraft pneumatic system was mentioned that weighed only one-half as much as an equivalent hydraulic system.

As might be expected, the matter of the weight of equivalent electric, hydraulic, and pneumatic systems for aircraft has received careful study, and the work of Geyer and Treseder¹ seems particularly complete. Their data on actual production components were for open-loop systems but may be applied directly to control systems also. The systems are broken into four categories: generation, storage, transmission, and actuator.

In Fig. 13.4 are shown the data on generation units. The production units for pneumatic systems are usually rated for less than 1 hp continuous

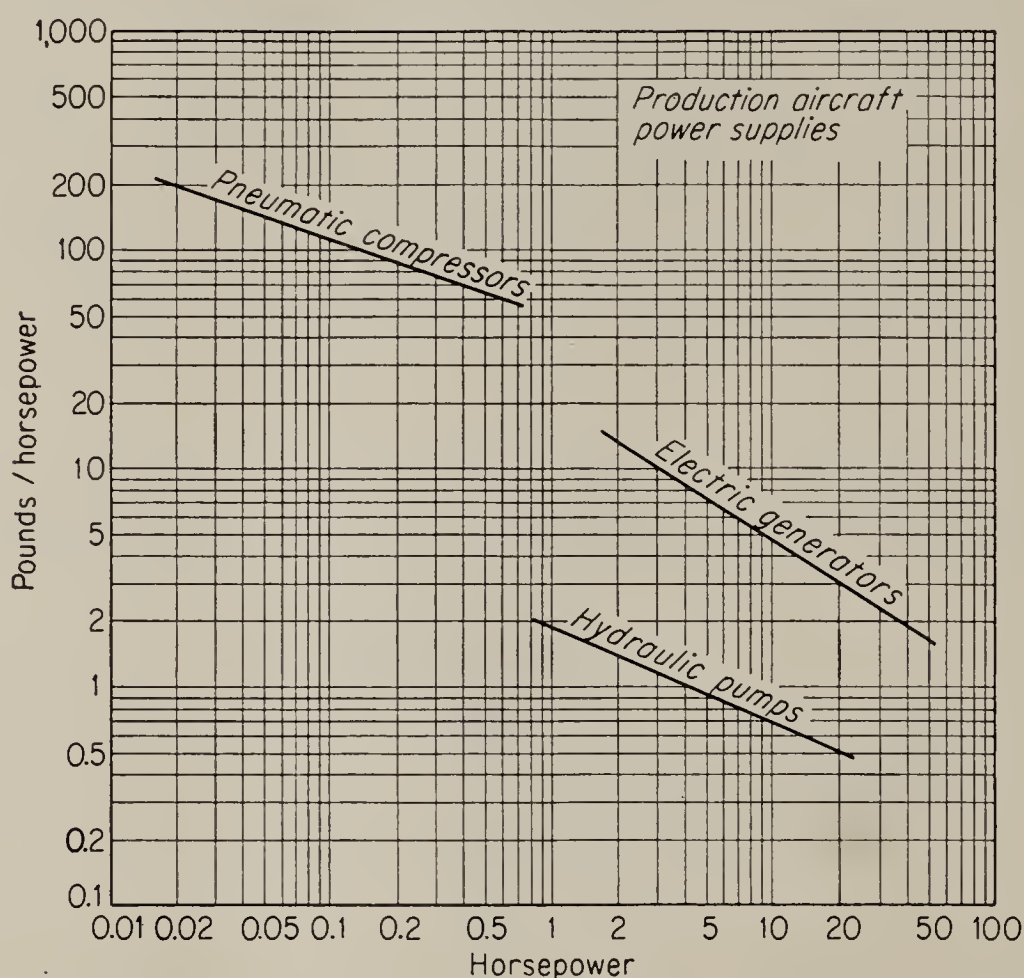


FIG. 13.4. Weight-per-horsepower comparison for modern aircraft power supplies. (Data from Geyer and Treseder)

duty, because storage of energy in pneumatic form is more economical, in terms of weight, than generation.

In Fig. 13.5 is shown the weight of energy-storage elements for the various types of systems. The weight of hydraulic accumulators and pneumatic storage cylinders is seen to be approximately equal, with the pneumatic having a slight edge. These figures include the weight of the working fluid. The data are plotted against the length of time that the storage device must deliver power to the load. It may be seen that the electric battery is the lightest storage form for loads of duration greater than 1 minute. The additional curves plotted in Fig. 13.5 for the weight of generating equipment allow a choice between generation and storage for loads for various durations. It may be seen that in practically

¹ H. M. Geyer and R. C. Treseder, Weight Analysis of Aircraft Actuators, *Trans. AIEE*, part II, vol. 71; pp. 118–126, 1952.

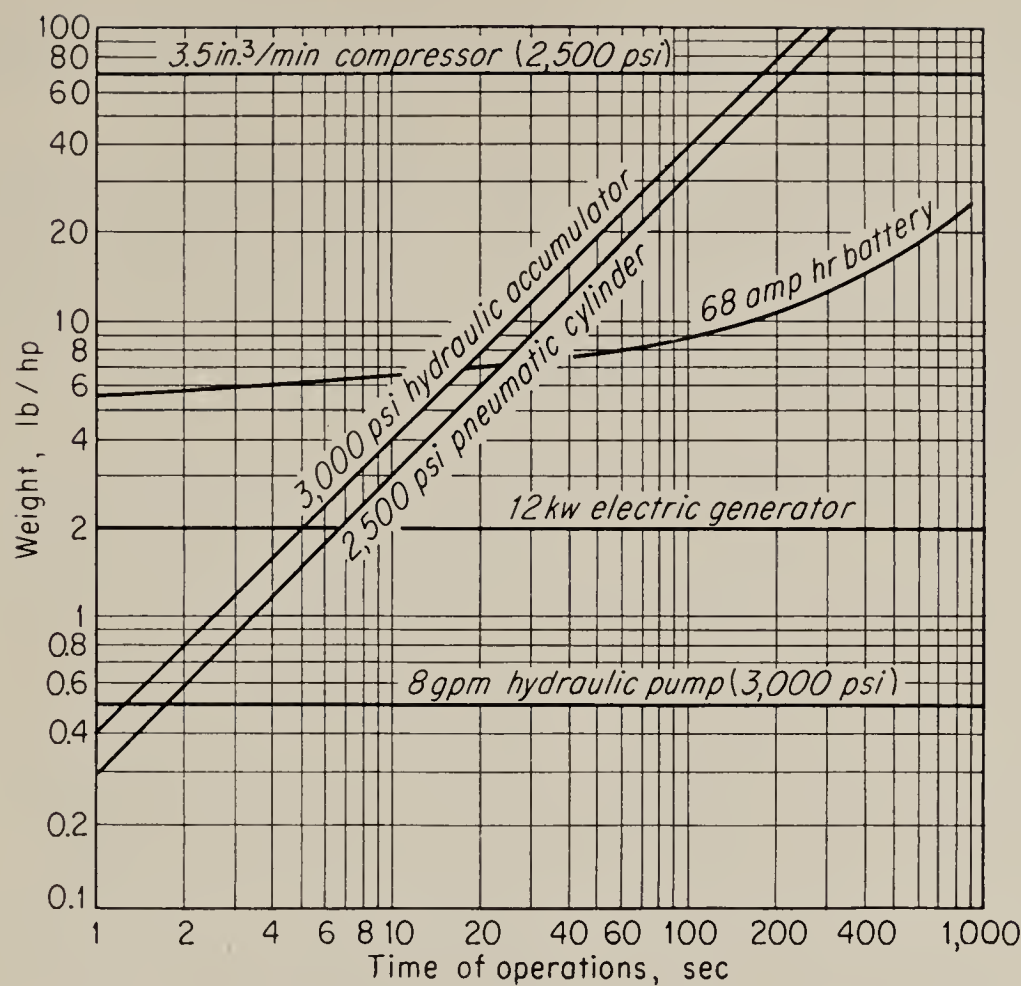


FIG. 13.5. Weight comparison for storage units as a function of time of operation. (Geyer and Treseder)

all cases it is more economical to provide adequate generating capacity in hydraulic and electric systems than to accumulate energy in a storage container. The hydraulic accumulator and the electric battery would be used only for smoothing shocks and for emergencies. In the case of the pneumatic system, however, the high weight cost of compressor horsepower renders the pneumatic storage element more economical for many loads. A major drawback, however, to the storage of large amounts of energy in pneumatic form is the danger of explosion should the accumulator be pierced.

It is in the transmission of power that the pneumatic system gains its major weight advantage over hydraulic systems. Figure 13.6 gives the weight cost of transmitting energy 75 ft. The pneumatic system is lighter because only one line is needed and the weight differential between the fluids is in its favor. The major drawback to the transmis-

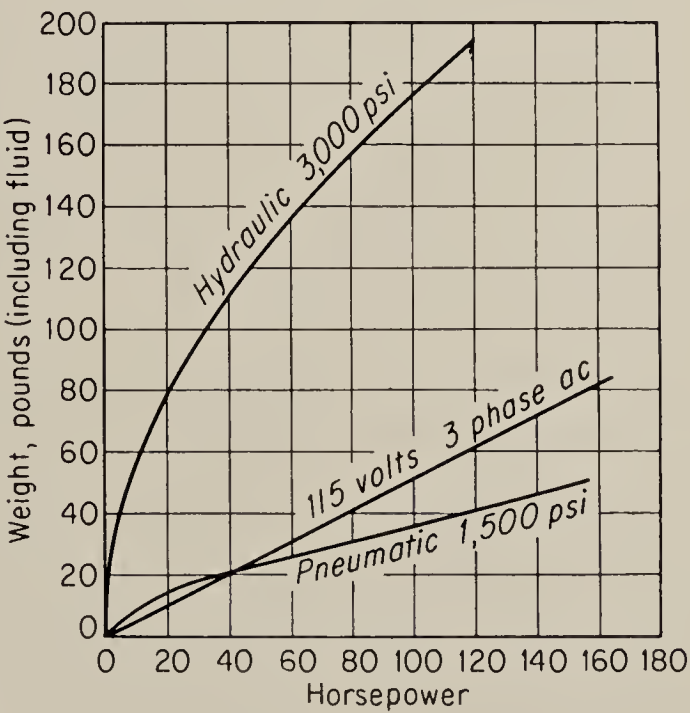


FIG. 13.6. Weight cost of the transmission of power 75 ft. (Geyer and Treseder)

Figure 13.6 gives the weight cost of transmitting energy 75 ft. The pneumatic system is lighter because only one line is needed and the weight differential between the fluids is in its favor. The major drawback to the transmis-

sion of power by compressed air over long distances is the transmission-line effects developed. The lags of pneumatic transmission lines were discussed in Chap. 12.

In Fig. 13.7 is shown a weight comparison of pneumatic, hydraulic, and electric actuators. It will be noted that, although the pneumatic actuator has a slight edge with respect to the hydraulic type, they can be considered equal for all practical purposes. Both are superior, it may be seen, to electric actuators.

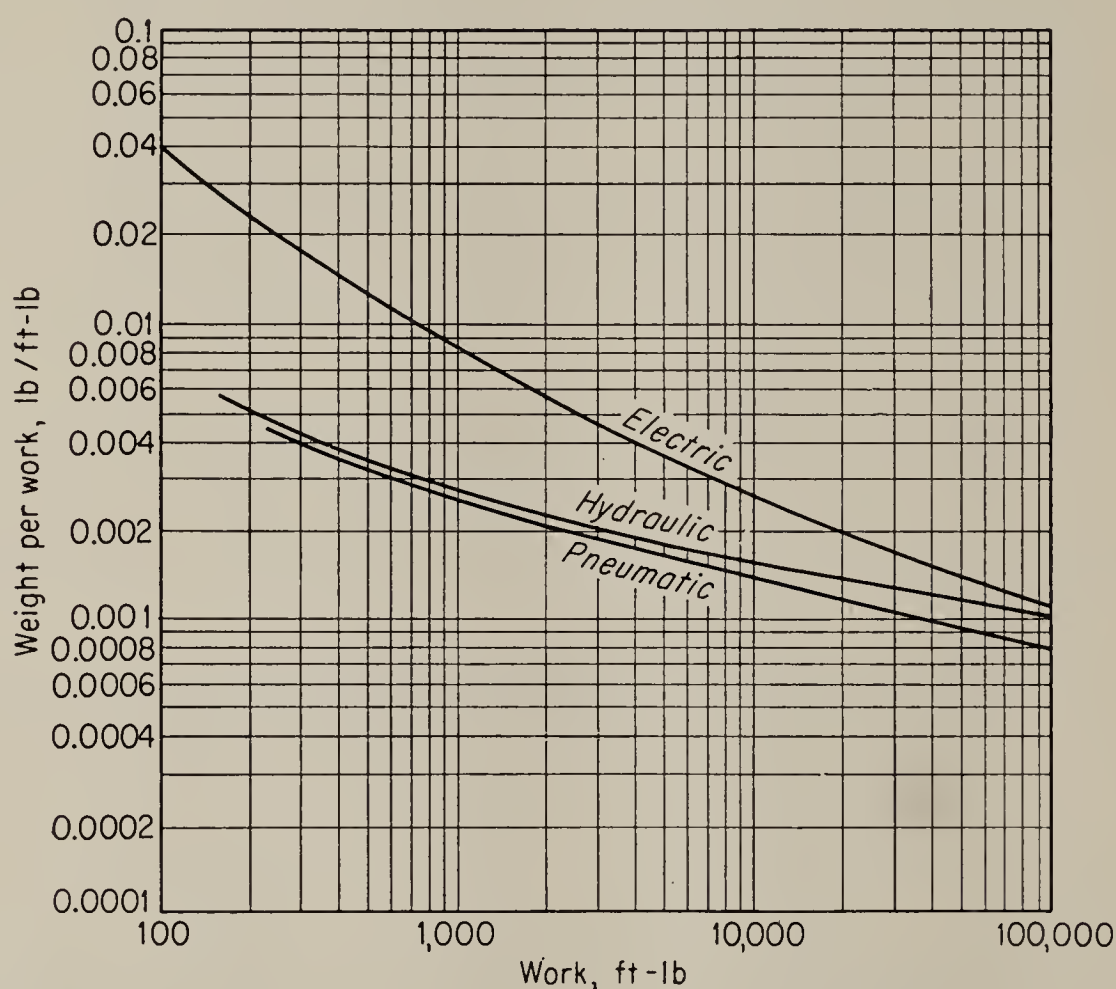


FIG. 13.7. Weight per unit of work versus rated work of actual production aircraft actuators. (*Geyer and Treseder*)

The pneumatic system, then, appears to be the most economical in weight in a moderately high peak horsepower system where a moderate to long transmission line is required and with a duty cycle that allows the energy to be stored in a large accumulator by a small compressor. This type of system is quite common in aircraft and guided missiles. In a system with other than these specifications, the pneumatic system might still be chosen to obtain one of its other advantages, as discussed in Chap. 12. The problem of establishing the optimum balance of the reserve energy supply for peak loads and the weight of stand-by equipment required to meet these relatively infrequent demands is one that must be solved for each individual case. It is conceivable that a system might have a pneumatic compressor capable of supplying less than one-tenth the peak load anticipated in a particular installation, if the peak

load were met infrequently. The accumulator could be charged initially from a fixed supply, and the system pump would act merely as a topping unit.

13.3. Choice of Operating Pressure. In the process-control industry, where size and weight are not of ultimate importance, the operating pressure of pneumatic systems is typically below 50 psig. The system components are then relatively large and easy to manufacture. Operation is reliable, and there is little danger of components under pressure exploding from rough handling. As the demands on speed of response are increased in the more critical processes envisioned in future developments, these operating pressures will be increased.

In high-speed pneumatic systems, especially air-borne systems, pressures have been continuously increased, although as yet there does not seem to be the standardization that there is in hydraulic installations. Aircraft pneumatic systems may operate from jet-engine compressor discharge pressure of 100 to 200 psi at sea level to about 50 psi at 50,000 ft altitude. Typical operating pressures for systems supplied by individual positive-displacement compressors are 500 psi, 1,000 psi, and 1,500 psi. The compressor charges an accumulator at about 3,000 psi, and the accumulator discharges to the system through a reducing valve. Experimental compressors and accumulators have been designed for 5,000-psi* storage for systems with 500- to 1,000-psi operating pressures. Schmidlin states that these 5,000-psi compressors are not appreciably heavier than the more usual 3,000-psi compressors and that the drive-power increase is only 15 per cent. The 5,000-psi accumulator is considerably smaller than an equivalent 3,000-psi element, the volume of the former being approximately 60 per cent of the volume of the lower-pressure element, as one would predict theoretically. Furthermore, the 5,000-psi accumulator requires walls of such a thickness that the container is essentially nonshatterable, whereas present design for the 3,000-psi unit requires wire-wrapped reinforcement. No doubt, both operating pressures and storage pressures will continue their present steady increase in the near future.

13.4. Pneumatic Power Supplies. The design of air compressors is beyond the scope of this volume, but the basic features necessary for intelligent use of compressors in control systems will be discussed. Pneumatic pumps, or compressors, may be of the turbo type or of the positive-displacement, or piston, type. For operating pressures above 50 psi the positive-displacement type is used almost exclusively because of its greater efficiency.

The p v diagram of an ideal reciprocating-compressor cycle is given in

* A. E. Schmidlin, Potential Advantages of a 5,000 psi Pneumatic System, *App. Hydraulics*, October, 1954.

Fig. 13.8. The intake stroke (1-2) is at inlet pressure. The compression stroke will lie between the extremes of perfect isothermal compression (2-3) and adiabatic compression (2-3'). The isothermal process requires that the heat due to compression be extracted continuously to maintain the gas at constant temperature, while the adiabatic process requires that

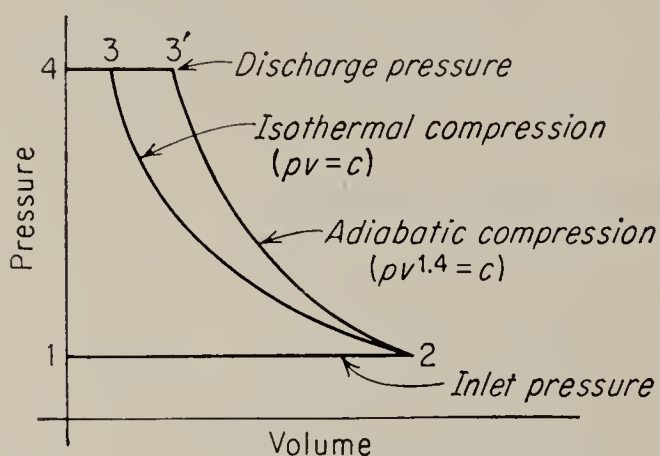


FIG. 13.8. The p v diagram of an ideal compression cycle.

the cylinder walls be perfectly insulated so that no heat escapes. The area enclosed by the cycle represents the work of compression; thus the isothermal cycle is to be preferred since it represents the smaller value of work. Even with isothermal compression, however, a good deal of energy is required to pump a volume of gas to an operating pressure of 800 or 1,000 psi. Let us say that an ideal pump is to compress air, at the

rate of 0.001 lb/sec to 1,000 psig. If perfect isothermal compression is assumed, we may use the familiar relation

$$\text{hp} = \frac{WRT}{550} \ln \frac{p_2}{p_1} \quad (13.2)$$

where W = weight of air, lb/sec

R = gas constant = 53.3 ft/° Rankine for air

T = absolute temperature, ° Rankine (530° Rankine = 70°F)

p = absolute pressure, psia

$$\text{Thus } \text{hp} = \frac{(0.001)(2.47 \times 10^5)(530)(6.94 - 2.71)}{(6,600)(386)} = 0.218 \quad (13.3)$$

To compress the same weight of air adiabatically requires

$$\text{hp} = \frac{WRT}{6,600g} \frac{k}{k-1} \left[\left(\frac{p_2}{p_1} \right)^{(k-1)/k} - 1 \right] \quad (13.4)$$

or

$$\text{hp} = \frac{(0.001)(2.47 \times 10^5)(530)}{(6,600)(386)} \frac{1.4}{0.4} \left[\left(\frac{1,015}{15} \right)^{0.286} - 1 \right] = 0.417 \quad (13.5)$$

In order to bring the efficiency of practical compressors as near the isothermal ideal as possible, several stages of compression are used, and interstage cooling is provided. Two or three stages are used with 1,000-psi units and as many as seven stages for 5,000-psi units. The cycle of a two-stage compressor is shown in Fig. 13.9. The intake stroke (1-2) remains the same as for a single-stage unit. The actual polytropic com-

pression stroke (2-5) for the first stage lies between the extremes of isothermal and adiabatic compression. Interstage cooling takes place at constant pressure (5-5'), and here it is assumed that the cooling is perfect. The second-stage compression stroke (5'-3'') then takes place. The shaded area represents the work saved by interstage cooling. Typical comparisons for work expended in one-, two-, and three-stage compression cycles are shown in Fig. 13.10.

In addition to improving efficiency, multistage compressors in which the compression ratio per stage is held down to a factor of 2 or 3, with interstage cooling provided, have another advantage. It is rather dangerous to compress even dry air with ratios of 8 or 10, since even a small amount of lubricating oil from bearings or oil working back from the

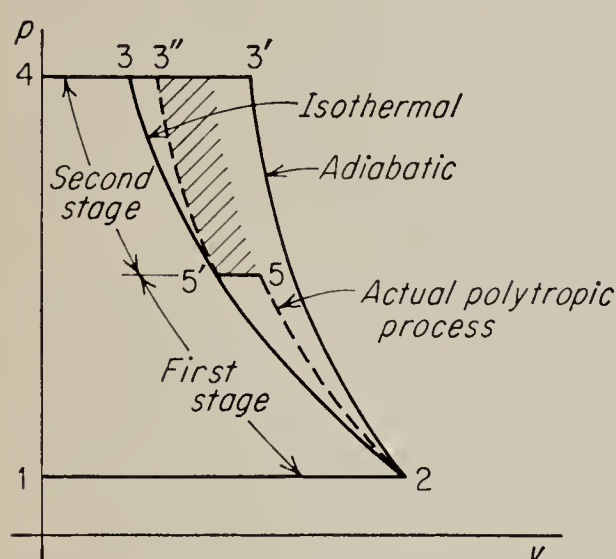


FIG. 13.9. The compression cycle of a two-stage compressor with interstage cooling.

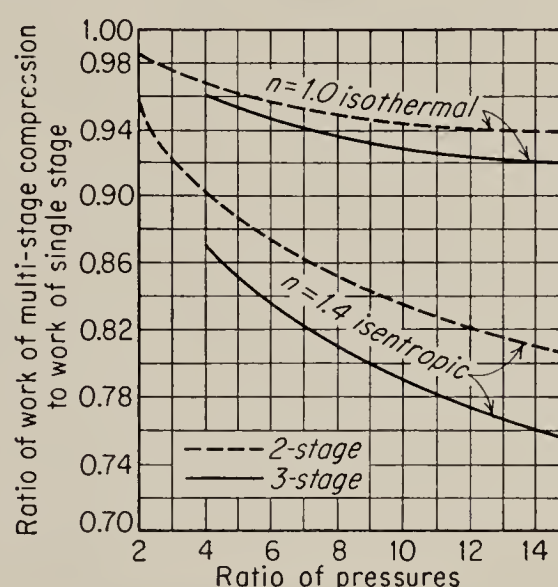


FIG. 13.10. Efficiencies of multistage compression cycles for pneumatic supplies. (From Lucke, "Engineering Thermodynamics," Columbia University Press, New York)

aspirator into the compression chamber may cause a diesel type of explosion at the high temperature and pressure in the chamber.

Much of the work actually done in compressing air cannot be recovered, because it is impossible to design nozzles and loads to allow the air to expand most efficiently, i.e., isothermally. Among other reasons this is due to the time that would be required for such an expansion. Thus over-all efficiencies of pneumatic systems are typically below 50 per cent, and figures as low as 30 per cent are sometimes quoted.¹ The efficiency of the compressor alone will usually lie between 70 and 85 per cent.

An exception to the general rule of positive-displacement compressors is found in jet aircraft, where the pneumatic supply may be the engine-compressor discharge pressure, as in the example discussed in Sec. 13.1. While this procedure is simple and convenient, it has at least two dis-

¹ J. L. Shearer and S. Y. Lee, Selecting Power Control Valves, *Control Eng.*, vol. 3, p. 72, March, 1956.

advantages. First, the compressor discharge pressure may vary by a factor of 5:1 from sea level to 50,000 ft, thus seriously affecting the performance of the pneumatic system. The second disadvantage occurs in pneumatic systems designed for relatively high power applications. Here the efficiency of the jet engine may be disturbed by heavy drains at the compressor discharge.

Even when the high-pressure air is supplied by a piston-type compressor, a change in altitude has an effect. A given compressor will not produce so large a weight of compressed air nor will its efficiency be constant as the ambient pressure is reduced.

13.5. Pneumatic Control Valves. The basic similarity between pneumatic valves and hydraulic valves allows us to base this discussion on Chap. 11 and, in the main, merely point out the differences between the devices. The basic configurations such as the piston valve and the flapper valve are quite the same for both liquids and gases. In fact the original development of the flapper valve seems to have been made for low-pressure pneumatic systems in the process-control industry and thereafter adapted for high-pressure hydraulic fluids.

Certain factors assume a different importance in pneumatic valves. The fluid reaction force, or Bernoulli force, for example, is negligible in the typical pneumatic control valve because of the very small mass rate of flow. For example, 1 ft³ of air at 1,000 psi weighs 0.0134 lb, whereas the weight of 1 ft³ of Univis J43 hydraulic fluid is 52.6 lb.

The viscosity of compressed air is negligible compared with that of hydraulic fluids, and thus the viscous damping in a pneumatic valve may usually be neglected. The low viscosity does, however, add one problem. The leakage flow in the clearances of the valve can be large. Even with clearances as small as a ten-thousandth of an inch the leakage flow is almost as large as it would be in an orifice of the same area as the clearance area;¹ i.e., the pressure drop in the leakage path is negligible. The characteristics of control valves can be calculated in the same manner as for hydraulic valves. Experience has shown that the flow process through such devices is very nearly adiabatic. Figure 13.11 shows the measured characteristic² of a nominally zero-lapped four-way slide valve of the type discussed in Chap. 11. The curves will be recognized as quite similar in form to the hydraulic-flow characteristics considered previously. Figure 13.12 is the calculated characteristic of an open-centered or under-lapped valve; again the curve will be recognized as similar to the corresponding hydraulic-flow characteristic.

The performance of a typical positioning system using air has been

¹ A. Egli, The Leakage of Gases through Narrow Channels, *Trans. ASME*, vol. 59, p. A53, 1937.

² J. L. Shearer, Study of Pneumatic Processes, *Trans. ASME*, vol. 78, p. 239, 1956.

compared to the performance of the same system using hydraulic fluid by Lee and Shearer.¹ A schematic diagram of the system is shown in Fig. 13.13. The supply pressure in both cases is 1,000 psi, and a slide valve is used to control the flow to a linear actuator. The quiescent

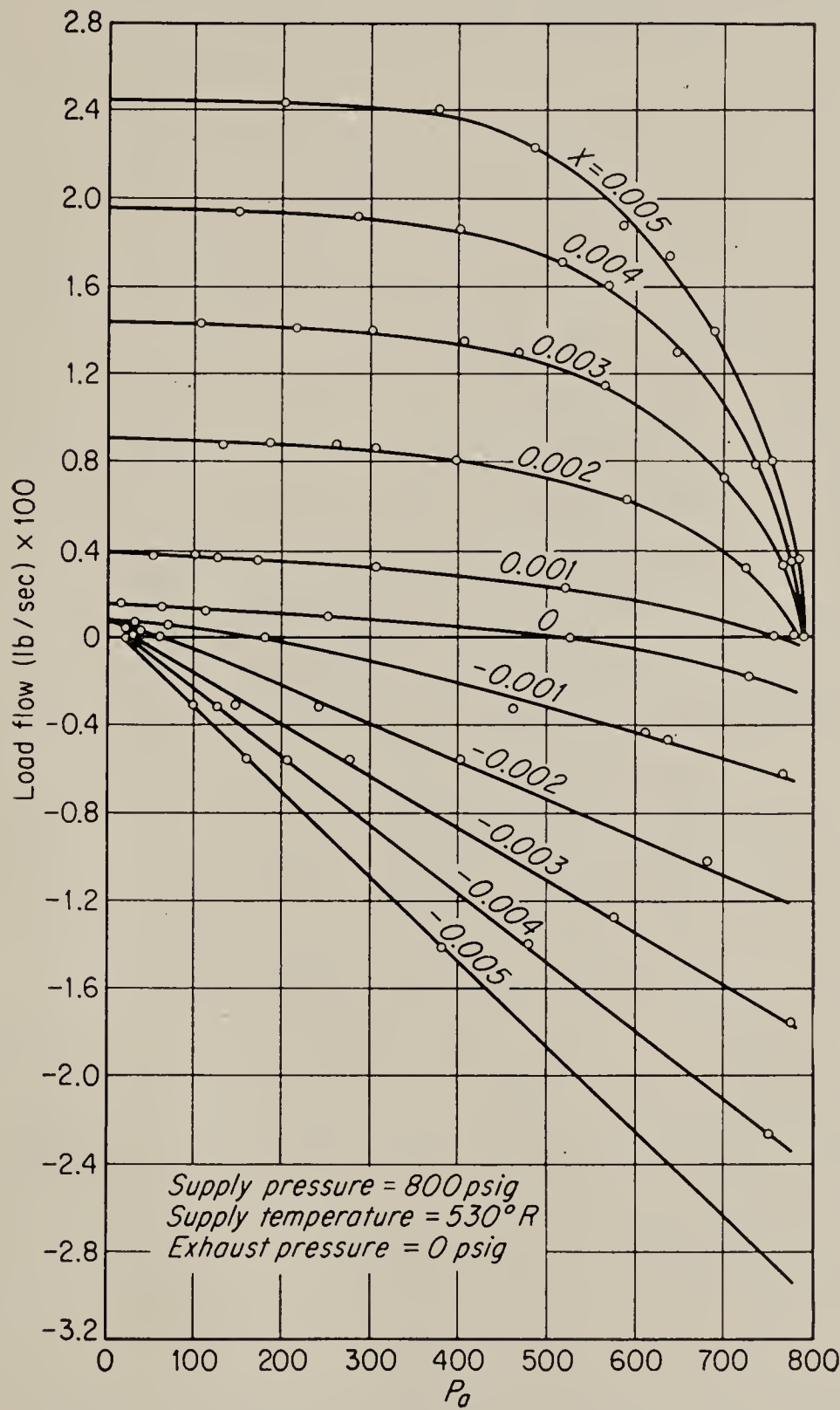


FIG. 13.11. Measured flow-pressure characteristic of a zero-lap valve as a function of stroke x . (Shearer)

leakage flow in the valve is 0.005 in.³/sec, and the maximum no-load velocity of the ram is 14 in./sec. In accordance with the principle discussed above, the exhaust port of the valve has a maximum diameter of 3/8 in., while the high-pressure port has a maximum width of only 1/4 in.

¹ S. Y. Lee and J. L. Shearer, Proceedings of the National Conference on Industrial Hydraulics, vol 8, p. 168, 1954.

The closed-loop frequency response of this servo is shown in Fig. 13.14, and for comparison is shown the frequency response of the same system when 1,000-psi hydraulic fluid is substituted for the compressed air. It is clear that the frequency response of the pneumatic system is inferior to that of the hydraulic one.

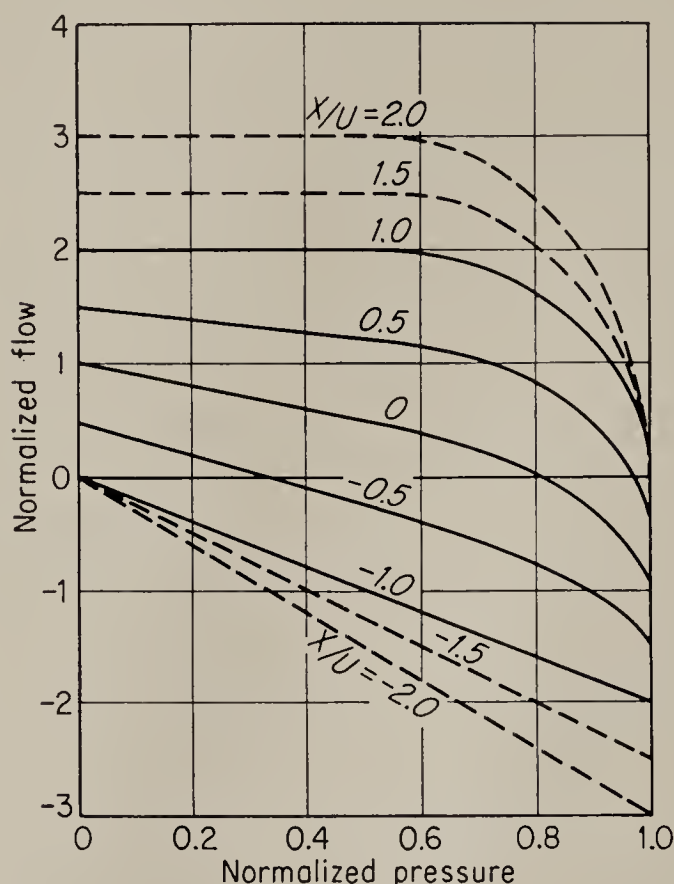


FIG. 13.12. Pressure-flow characteristic of an underlapped pneumatic valve with the ratio of stroke x to underlap u as a parameter. (Shearer)

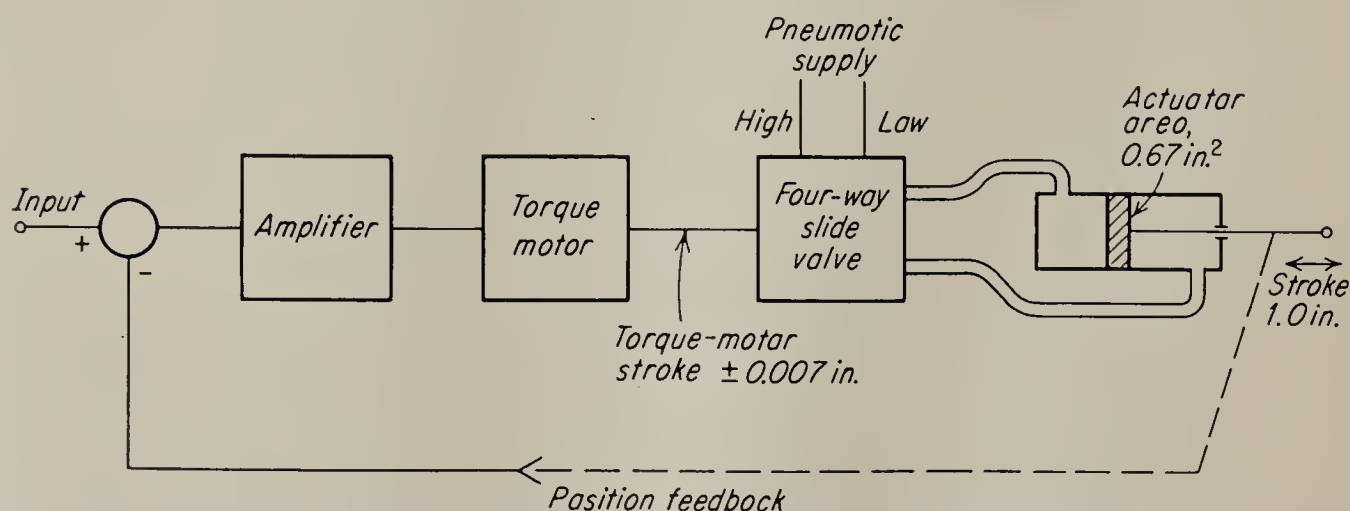


FIG. 13.13. Diagram of a pneumatic positioning system.

Many pneumatic positioning systems are designed with spring-loaded actuators that may take the form of a bellows or a Bourdon tube. In such a case the valve should control pressure rather than flow. Of course, it is rather artificial to make this distinction between pressure control and flow control, since any valve is merely an adjustable orifice that actually controls only the relation between pressure and flow, but we are conforming here to common usage. An open-center slide valve or spool valve

may be considered as a pressure-control valve in this sense, as may a flapper valve. In each case it is not necessary that there be load flow in order to establish a required load pressure. The load pressure is established by positioning the valve in the underlap region so that the leakage flow to sump sets up the required pressure drop in the valve.

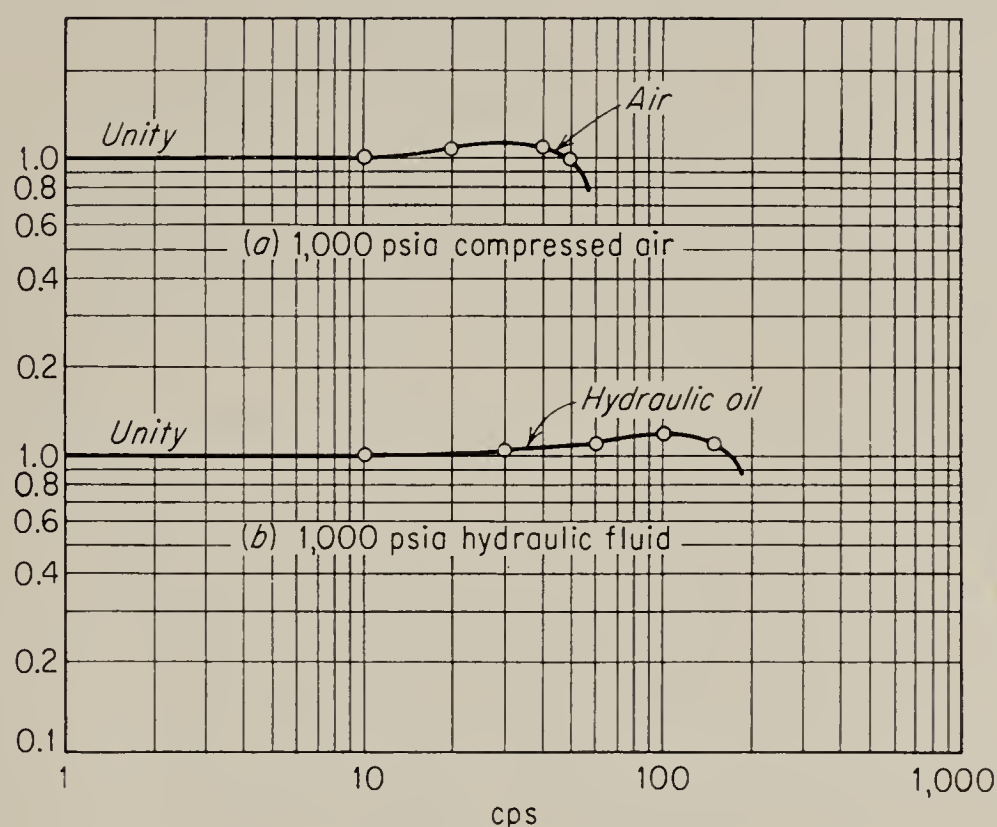


FIG. 13.14. The frequency response of the positioning system shown in Fig. 13.13. (From Lee and Shearer)

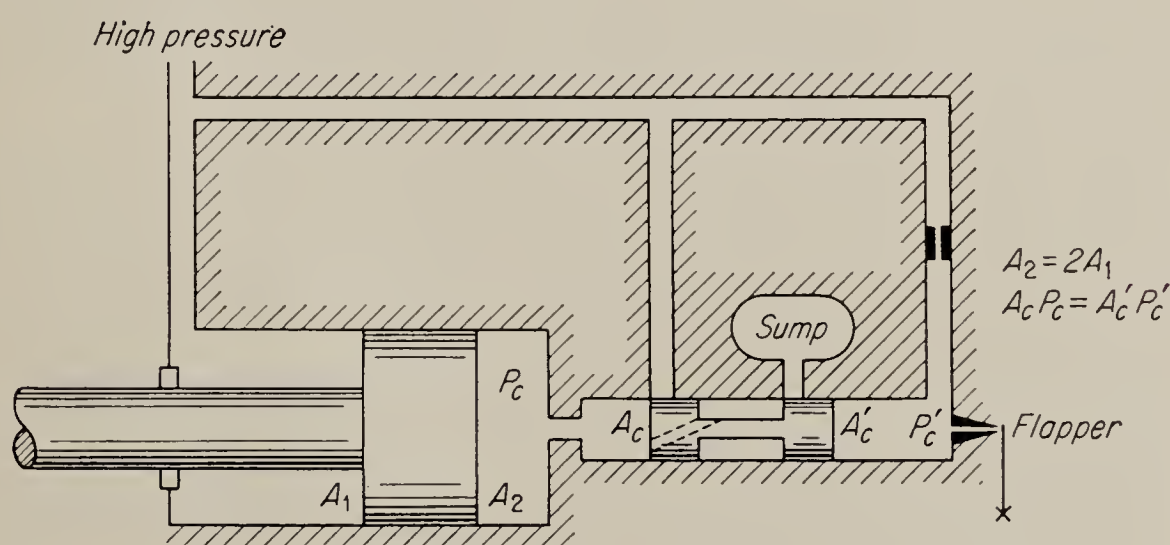


FIG. 13.15. A two-stage, flapper-controlled pressure valve with differential-area actuator.

Two pressure-control valves, somewhat more complex, are shown in Figs. 13.15 and 13.16. The valve in Fig. 13.15 has a free piston that is positioned by the pressures on its two ends. P'_c is controlled by the flapper stage, and the spool will move to such a position that the flow from the supply through the high-pressure orifice and then the low-pressure orifice to sump will make P_c such that

$$P_c A_c = P'_c A'_c$$

Since P_c is also applied to one end of the differential-area actuator, a change in P_c results in an acceleration of the unloaded actuator. So long as the size of the chambers is small enough to be negligible, the amplitude ratio varies inversely with the second power of frequency. If the differential-area actuator is replaced by a bellows, the reader may show that the position of the flapper controls the position at the output of the bellows.

The valve shown in Fig. 13.16 is of the so-called nonbleed type and is widely used in the process-control industry. The bellows limit its use to the lower range of operating pressures. The valve has two stages, the first consisting of the input flapper f_1 and nozzle n_1 arrangement that

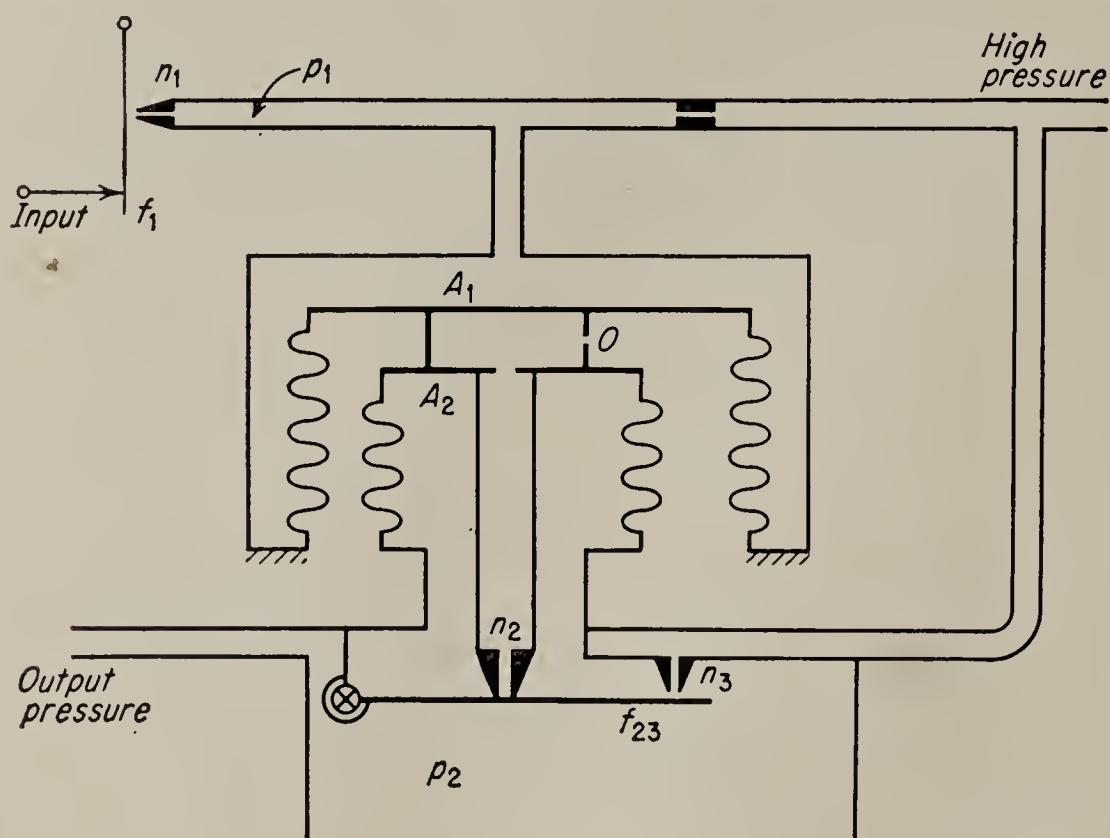


FIG. 13.16. Another type of two-stage, flapper-controlled pressure valve, called a nonbleed relay in the process-control industry.

controls the pressure p_1 . The second stage consists of the dual bellows and the two nozzles n_2 and n_3 with their common flapper f_{23} .

In equilibrium, $p_1 A_1 = p_2 A_2$, n_2 is closed and n_3 is almost closed by flapper f_{23} . If p_1 increases, the net bellows force carries n_2 down, thus forcing f_{23} away from n_3 and allowing p_2 to increase, returning n_2 and n_3 to equilibrium. If p_1 decreases, n_2 moves up, leaving f_{23} to close n_3 and allowing p_2 to vent to atmosphere through orifice O until the pressure ratio is reestablished. The “nonbleed” appellation can be perhaps justified since bleeding only takes place when either p_1 or p_2 varies.

13.6. Pneumatic Motors. Rotary motors of any of the types discussed under hydraulic components should theoretically operate properly under pneumatic pressure, provided only that they receive proper lubrication. At the present time, however, the only rotary-motor type commercially

available for high-pressure pneumatic applications is of the vane type. The clearance volume of this device, shown in Fig. 13.17, is larger than the volume of a hydraulic motor of the same horsepower rating, but the operating principle is identical.

The high-pressure fluid is admitted at the inlet port and acts against the differential area of the vane, developing a torque. The rotor turns and carries the vane past the outlet port, where the air is expelled to atmosphere. The motor must be symmetric for reversible operation, and no energy is extracted from the expanding air, thus wasting the energy of compression. Although the vanes may be spring-loaded, the typical vane motor depends on centrifugal force to keep the vane seated against the housing. The spring-loaded vane is advised when the motor

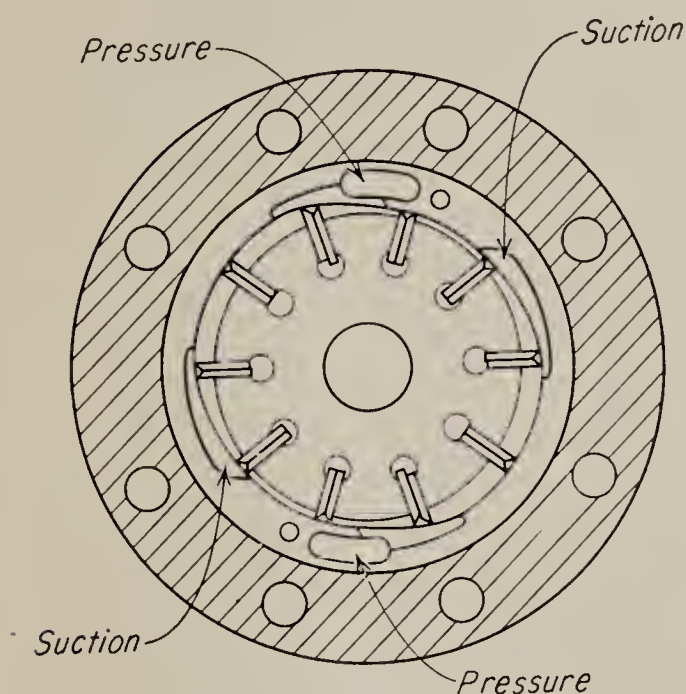


FIG. 13.17. Vane-type pneumatic motor.

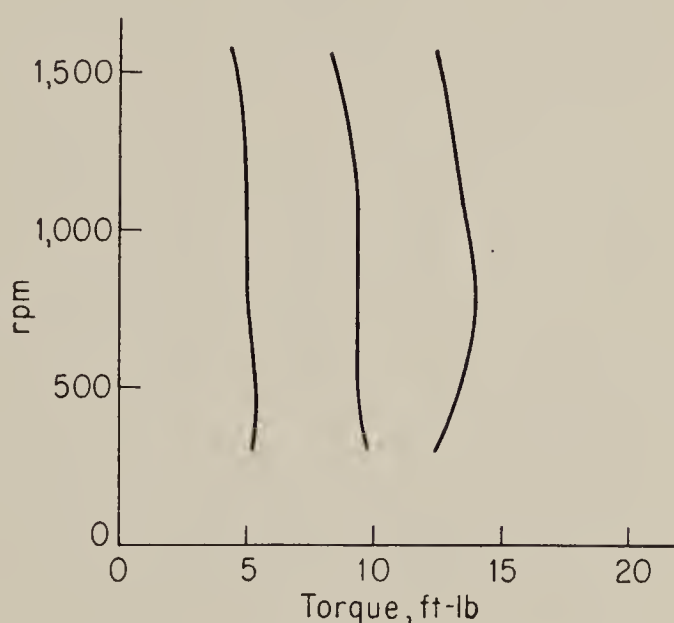


FIG. 13.18. Speed-torque curves with operating pressure a parameter for a pneumatic vane motor.

is to be stalled in normal operation. Typical speed-torque curves for a vane motor as a function of operating pressure are given in Fig. 13.18. Since torque is proportional to pressure in pneumatic devices as in hydraulic devices, the ideal characteristic would be a straight vertical line. It may be seen that actual device corresponds closely to theory.

Up to the present time the gear-type design has not been successful either as a hydraulic or a pneumatic motor, owing to the high starting friction that is usually involved. It appears, however, that much of this friction is due to improper design arising from attempts to produce an economical unit. A properly engineered and manufactured gear motor can be made to have as little as 5 lb-in. of static friction and to break free and run at as little as 20-psi pneumatic pressure. These units will operate successfully as pneumatic motors if the compressed air is supplied with an oil spray from an *aspirator* in the line. The oil suspended in the com-

pressed air not only acts as a lubricant but also allows an oil film to be built up which seals the meshed gears and contributes to the volumetric efficiency of the device by preventing leakage between the gears to sump. Figure 13.19 shows speed-torque curves for a gear motor designed for pneumatic service.

The piston-type hydraulic motor should be capable of operation as a pneumatic motor, but a curious phenomenon is sometimes observed when this application is attempted. After a short period of proper operation, a sort of diesel action begins to take place. Apparently, the compressed oil vapor suspended in the air begins to explode behind the cylinders. At

the same time, however, the motor is refrigerated by adiabatic expansion of the air in the chambers, and the motor case remains cool. This diesel action probably reduces the efficiency of the device and may damage the mechanism if allowed to continue.

13.7. Actuators. Linear actuators of the piston type used for hydraulic systems are used also for compressed-air service. Both spring-loaded and differential-area types are also in use. The only construction difference appears in the packing and seals. The spring-loaded actuator seems more common in pneumatic systems, and it may take the form of a bellows or a Bourdon tube as well as the piston type.

Bellows and Bourdon tubes are also useful as temperature-sensitive and pressure-sensitive devices in both hydraulic and pneumatic control

systems. When these elements serve as temperature-sensitive devices, their principle of operation depends on the temperature coefficient of expansion of the fluid enclosed. In the pressure-sensitive application, the linear extension of the device depends on the pressure of the fluid enclosed.

As a spring-loaded linear actuator, the Bourdon tube usually has much the stiffer spring constant. While it is possible to design a bellows with a very high spring rate for applications where only a small extension is necessary, the Bourdon tube will usually exhibit better repeatability in position versus pressure and will have the longer life. Bourdon tubes

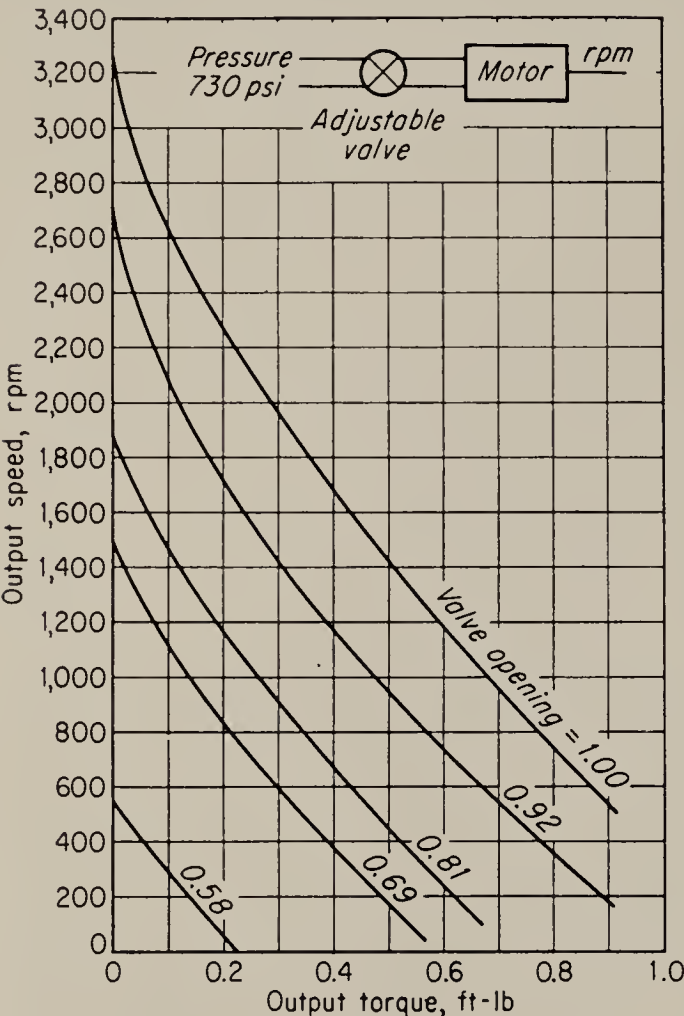
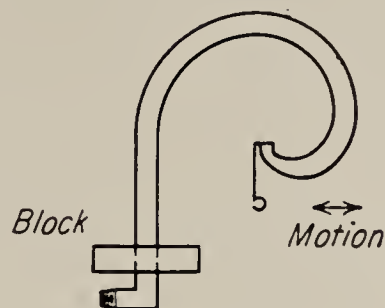


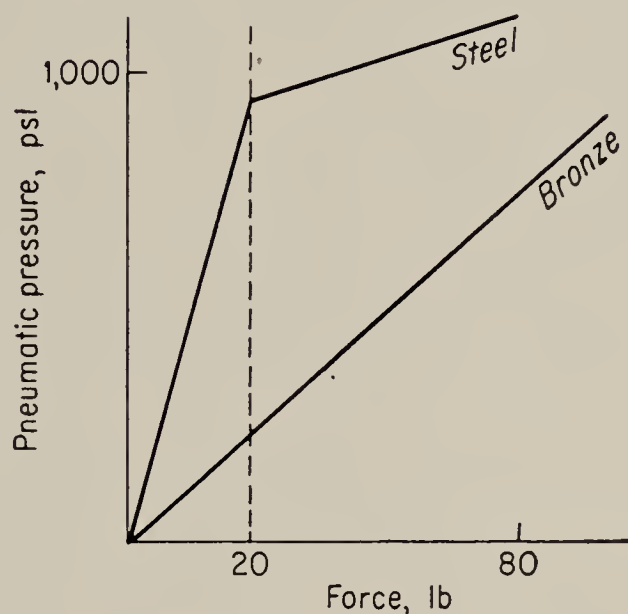
FIG. 13.19. Speed-torque curves for a pneumatic gear motor.

are constructed of steel or brass and typically exhibit excellent linearity within their operating range. Typical characteristics are given in Fig. 13.20. Neither Bourdon tubes nor bellows should be allowed to expand beyond their free length if long life is to be assured.

While bellows have been constructed of many types of materials, the metal bellows must be used in high-performance, high-pressure applications. Metal bellows may be employed in systems with pressures up to and beyond 3,000 psi. Bellows are used in many aircraft control systems as pressure-sensitive devices that correct for altitude, e.g., in bombsights and fuel controls to set the fuel-air ratio. Bellows can be constructed to sense differential pressure, as shown in Fig. 13.21. The pressures p_1 and p_2 are introduced to the inside and the outside, respectively, of the large bellows. The large expanding bellows is sealed at both ends by the two small bellows contained inside it. The motion of the actuating rod is then a function of the pressure difference. The unit is more compact and has a higher sensitivity than two separate bellows connected by mechanical linkage.



(a) Schematic



(b) Typical characteristics

FIG. 13.20. Bourdon tubes: (a) schematic; (b) typical characteristics.

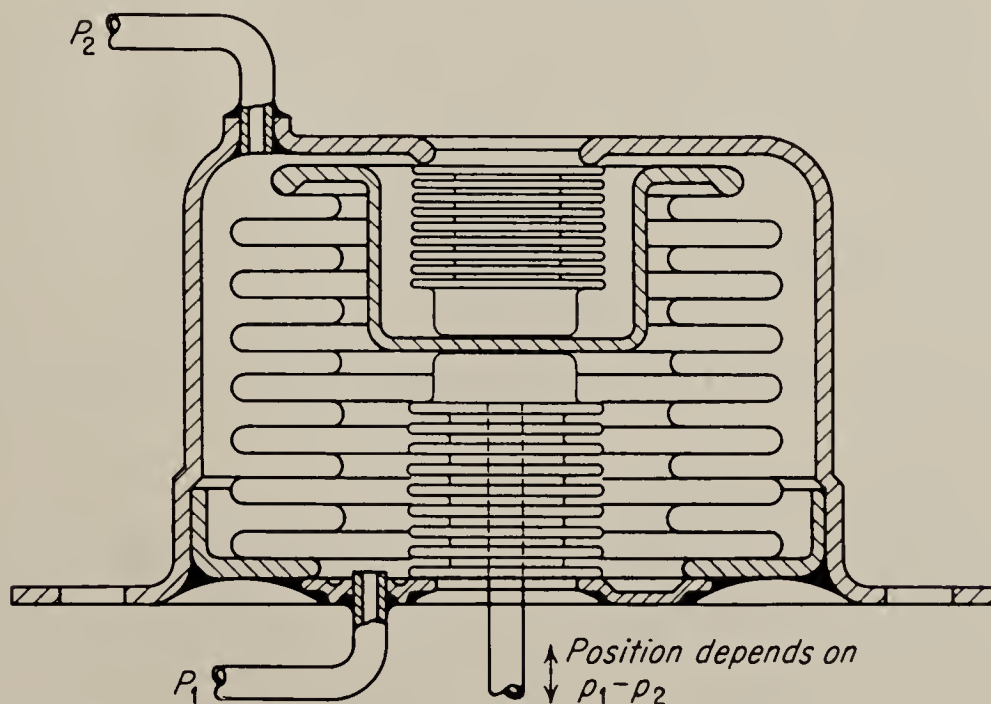


FIG. 13.21. Differential-pressure bellows. (Howard)

Various metals are used in bellows construction depending on the service factors important in the particular application. Table 13.1 summarizes present experience with the various common materials. Bellows are constructed either with standard or extra-flexible convolutions. The standard convolutions have hysteresis values ranging from 4 per cent for

TABLE 13.1. METALS USED IN BELLOWS CONSTRUCTION*

Metal	Corrosion resistance	Max temp, °F	Ease of fabrication	Cost
Brass.....	Fair	350	Excellent	1
Brass, silver-clad†.....	Excellent	350	Excellent	3
Bronze.....	Fair	350	Excellent	2
Monel.....	Good	900	Fair	4
Stainless steel.....	Excellent	1100	Fair	5
Inconel.....	Excellent	1500	Fair	6 (high)

* J. H. Howard, Metal Bellows, *Machine Design*, vol. 26, no. 1, pp. 137–148, 1954.
† The silver-clad brass is not plated, but rather the solid silver is rolled on.

TABLE 13.2. CHARACTERISTICS OF 1½-IN. BELLOWS*

Characteristic	Regular				Extra-flexible		
Root diameter,† in.....	¾				2⅓₂		
Convolutions (standard)‡.....	11				11		
Effective area,§ in.².....	0.69				0.62		
Wall thickness, in.....	0.004	0.005	0.006	0.007	0.004	0.005	0.006
Approx. free length, per convolution, in.....	0.084	0.079	0.074	0.070	0.100	0.087	0.076
Maximum deflection,¶ per convolution, in.....	0.029	0.028	0.027	0.023	0.047	0.034	0.029
Spring rate,** lb/in. per convolution	172	291	552	687	62	108	226
Flexibility,†† psi per convolution, in/psi.....	0.0040	0.0024	0.0012	0.0010	0.010	0.006	0.003
Maximum internal pressure, psi.....	55	80	150	200	55	65	130
Maximum external pressure, psi....	60	88	165	220	61	72	143

* J. H. Howard, Metal Bellows, *Machine Design*, vol. 26, no. 1, pp. 137–148, 1954.
† To obtain approximate inside diameter, subtract two times the wall thickness from root diameter.
‡ May be decreased in every case, increased in some.
§ To obtain volume, multiply by bellows length.
|| To obtain total free length of bellows, multiply by number of convolutions.
¶ To obtain total maximum deflection of bellows, multiply by number of convolutions.
** To obtain spring rate of bellows, divide by number of convolutions.
†† To obtain flexibility of bellows, multiply by number of convolutions.

brass to 1.0 per cent for bronze. These figures are reduced by 2 by heat-treating and by a factor of about 3 for extra-flexible convolutions.¹

Design tables are available from manufacturers for the physical dimensions of all standard bellows. Table 13.2 is an example of one standard table for 1 1/8-in. convolutions.

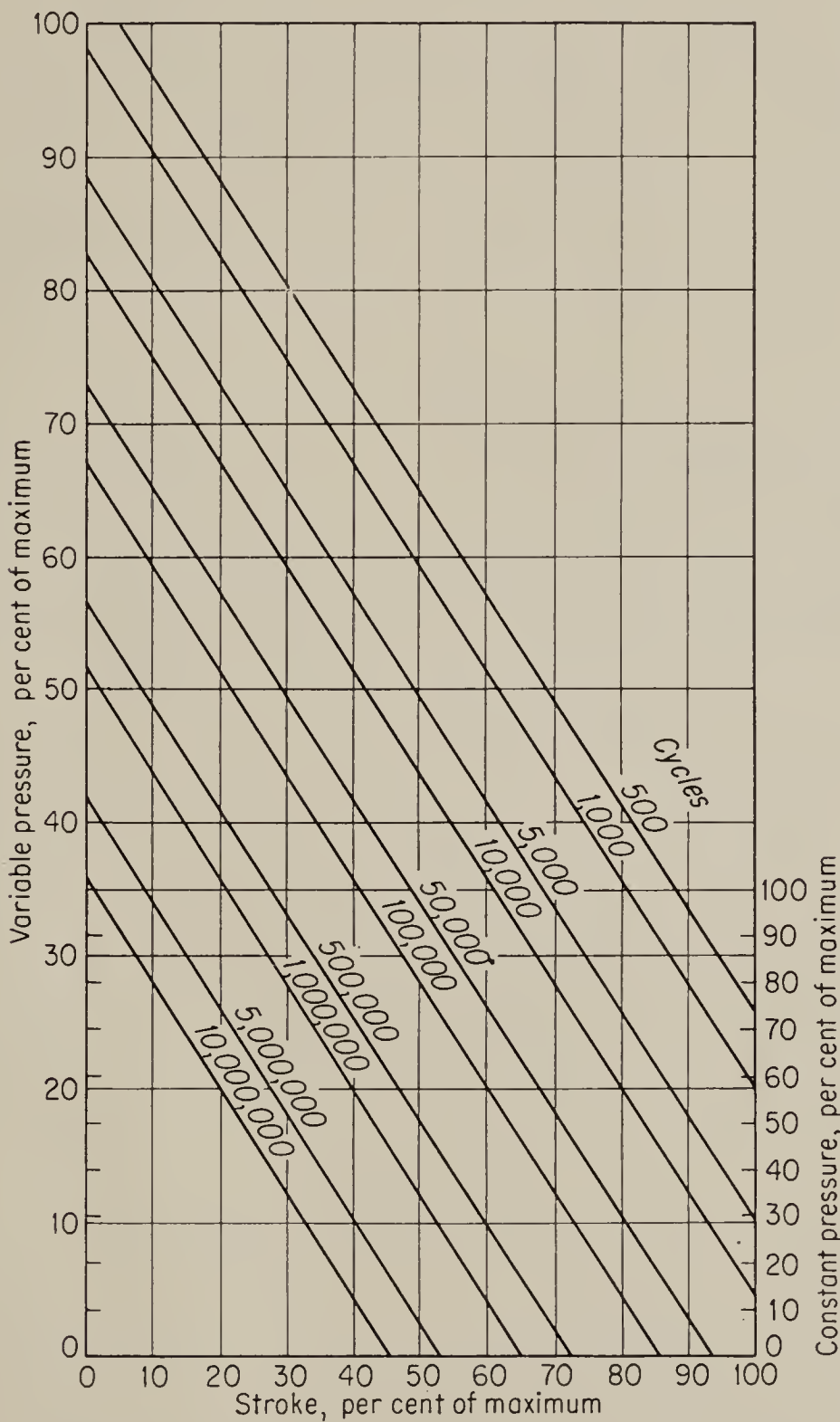


FIG. 13.22. Life expectancy of metal bellows as a function of pressure variation and stroke. (Howard)

When bellows are subjected to large amounts of flexing or large variations in pressure, their life expectancy is reduced. This factor must be considered in the design. Figure 13.22 shows the life expectancy of metal bellows in cycles as a function of pressure variation and per cent stroke.

¹ J. H. Howard, Metal Bellows, *Machine Design*, vol. 26, no. 1, pp. 137–148, 1954.

Bellows should never be expanded beyond their free length. The figure for variable-pressure life must be used when pressure changes, even if the bellows remains fixed. Internal flexing takes place as the pressure changes, even if the ends are fixed, thus reducing the life. Several common bellows configurations are shown in Fig. 13.23.

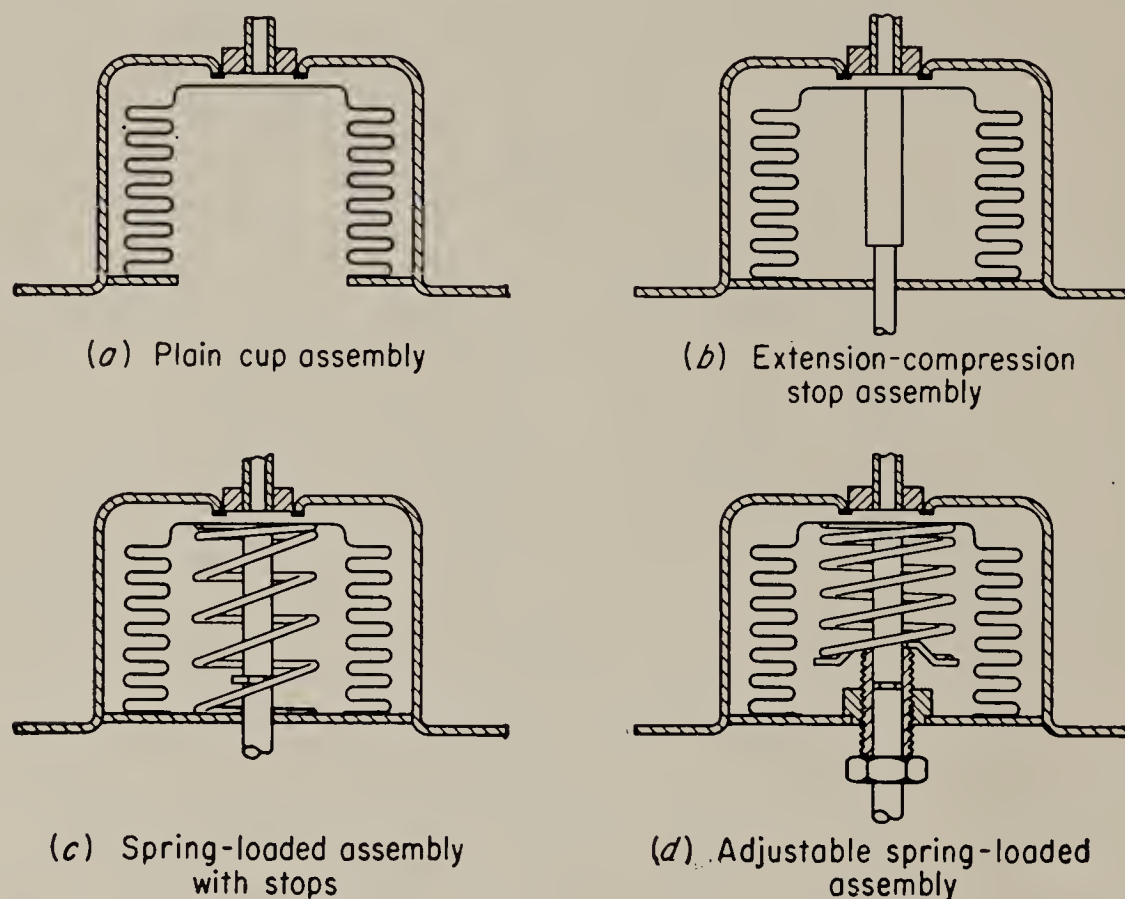


FIG. 13.23. Common types of bellows assemblies. (Howard)

Howard¹ sums up the properties of bellows as follows: flexibility, or inches of stroke per pound per square inch of pressure, varies directly with the number of convolutions, as the square of the outside diameter, inversely with the cube of wall thickness, and inversely with the modulus of elasticity of the bellows material. Spring rate, of course, is the inverse of flexibility.

PROBLEMS

13.1. Calculate the input impedance of the transmission line and load in Prob. 10.1 for compressed air as a fluid.

13.2. Discuss the advantages and disadvantages of using compressed nitrogen as a fluid in a pneumatic system.

13.3. A missile manufacturer is investigating the use of high-pressure superheated steam as the driving fluid in a pneumatic system. The fluid will be stored as water and passed along inside the skin of the missile for cooling, thus being converted to steam. Discuss the practicality and implications of such a proposal.

¹ *Ibid.*

NAME INDEX

Adler, R., 1
Ahrendt, W. R., 150, 238
Aikman, A. R., 447, 458, 459
Alexanderson, E. F. W., 198
Anner, G. E., 182, 380, 456

Barnes, J. L., 11, 270, 442
Beam, A., 331
Berard, S. J., 316, 319
Beyer, G. L., Jr., 229
Binder, R. C., 437, 439
Blackburn, J. F., 397, 399
Bode, H. W., 18, 20, 196
Bower, J. L., 21, 22, 34, 200, 251, 418
Bowman, K. K., 198
Bradner, M., 442
Bright, R. L., 272, 273
Brown, L. O., Jr., 289
Bull, H. S., 180
Burnett, J. H., 149
Burns, L., Jr., 1

Cahn, S. L., 420
Caldwell, W. I., 447
Carr, C. C., 8
Chang, S. S. L., 302
Chestnut, H., 11, 13, 16, 43, 83, 229, 230, 328, 417
Chin, P. T., 130
Chubbuck, J. G., 389
Conrad, A. G., 229, 279
Cook, A. L., 8
Cooke, C., 431
Corcoran, G. F., 228
Cunningham, W. J., 407, 408

Davison, L. M., 317
Dawes, C. L., 180, 187, 208
Den Hartog, J. P., 346
Dodge, R. A., 377, 392, 404
Doetsch, G., 11
Dornhoefer, W. J., 173

Dow, W. G., 119, 150
Draper, C. S., 220, 355, 356
Dushkes, S. Z., 420

Edwards, M. A., 198
Egli, A., 470
Everitt, W. L., 182, 186, 380, 456

Famme, J. H., 436
Firestone, F. A., 306
Fleck, J. T., 22, 34
Formhals, W. H., 192
Frazier, R. H., 298
Fried, D. D., 229, 244

Galonska, D. A., 333
Gardner, M. F., 11, 270, 442
Gerwing, H. F., 436
Geyer, H. M., 464
Geyger, W. A., 165, 177
Gibson, J. E., 150, 308, 389
Goldberg, E. A., 102
Greening, C. P., 361
Greenwood, I. A., Jr., 150, 259, 315
Greer, E. M., 431
Grohe, L. R., 356
Guillemin, E., 18, 20, 37, 88

Hadekel, R., 424
Helm, H. A., 456
Hetenyi, M. I., 339
Hilbourne, R. A., 120
Holdam, J. V., Jr., 150, 259, 315
Howard, J. H., 478
Hunter, L. P., 87-89, 94, 95, 97

Jackson, K. R., 389
James, H. M., 16, 43, 238, 265, 266, 320
Janssen, J., 454, 460
Jones, D. D., 120

- Kerchner, R. M., 228
Kochenburger, R. J., 152, 216
Koopman, R. J. W., 284, 294, 295, 299
Korn, G. R., 6
Kretzmer, E. R., 95
Kron, G., 302
Kronacher, G., 229
Krummenacher, V. H., 173
Krupe, A. P., 272, 273
- Langsdorf, A. S., 242
Lee, S. Y., 397, 399, 421, 422, 437, 469, 471
Lees, S., 220
Levenstein, H., 441
Lloyd, T. C., 229, 279
Lyon, W. V., 284
- McDonald, D., 313
McKay, W., 220
MacRae, D., Jr., 150, 259, 315
Mapes, T., 389
Martin, L. D., 317
Mayer, R. W., 11, 13, 16, 43, 83, 328, 417
Michalec, G. W., 330
Miller, S. E., 74
Moyer, E. E., 130
Murphy, R. J., 371
- Newton, G. C., 369
Nichols, N. B., 16, 43, 238, 265, 266, 320
Nightingale, J. M., 419
- Oldenburger, R., 456
Oliver, B. M., 391
- Page, L., 349
Parker, N. F., 361
Pestarini, J. M., 119
Peterson, A., 124
Peterson, B., 325, 326
Phelps, C. W., 316, 319
Phillips, R. S., 16, 43, 238, 265, 266, 320
Pierce, J. R., 391
Preston, S. T., 345
Puchstein, A. F., 229, 279
- Ramey, R. A., 175
Ramo, S., 10
- Rathbone, E. H., 318
Rawlings, A. L., 352
Reed, W. E., 461
Reich, H. J., 68, 81, 115, 128, 150
Richardson, K. I. T., 352
Roberts, W., 1
Rosenbloom, J. H., 229, 243, 244
Russell, W. T., 361
Rutherford, C. I., 447, 458, 459
- Schmidlin, A. E., 467
Schuler, M., 352, 360
Scrafford, R. L., 425
Shannon, C. E., 391
Shea, R., 88, 94, 95, 121
Shearer, J. L., 422, 437, 469, 470, 471
Sinclair, D., 124
Skalnik, J. G., 129
Slaughter, D. W., 95
Soroka, W. W., 334
Stanton, L., 43
Stenning, A. H., 439
Stolarik, E., 444
Storm, H. F., 161, 177
Sverdrup, N. M., 377, 394
Svoboda, A., 334, 336
Sweeney, D. C., 404, 405
Sziklai, G. C., 126
- Thompson, M. J., 377, 392
Treseder, R. C., 464
Trickey, P. H., 302
Trout, W., 371
Truxal, J., 120, 148, 268-270, 308, 312
Tustin, A., 447, 454
Tuteur, F. D., 150, 389, 418
- Valley, G. E., 43, 52, 53, 115, 254
Vander Kaay, H. A., 371
Von Mises, R., 398
- Wallman, H., 43, 52, 53, 115, 254
Waters, E. O., 316, 319
Webb, R. C., 440
Weiss, G. H., 229-231, 233, 243
Whinnery, J. R., 10
Williamson, H., 447
Wrigley, W., 356
- Zemmerli, R., 327
Zweig, F., 370, 380, 382

SUBJECT INDEX

- A-c servo (*see* Carrier-frequency servo; Servomotors)
- Accelerometer, 357–359
- Accumulator, 430
 - pneumatic, 464, 467
- Accuracy, 200, 221
- Actuator, hydraulic, 386
 - pneumatic, 476–480
- Air compressors, 467–470
- Ambiguity in speed synchro systems, 240–241
- American Gear Manufacturers Association, 317, 328
- Amplidyne, 180, 198–206
 - analysis, 200
 - inaccuracies in, 206
 - armature reaction in, 187, 199
 - compensating winding, 187
 - deadband effect, 206
 - definition, 198
 - direct-axis time constant, 205
 - direct-axis voltage, 199
 - effect of nonlinear brush resistance, 206
 - effectiveness of direct-axis compensating winding, 201
 - equivalent circuit, 205
 - field time constant, 205
 - hunting, 206
 - mutual inductance, 188, 200
 - open-circuit voltage expression, 203
 - output-impedance expression, 204
 - principle of operation, 198
 - quadrature-axis brushes, sparking at, 200
 - quadrature-axis time constant, 205
 - quadrature-axis voltage, 199
 - speed voltage parameter, 200
 - transfer function, low-frequency approximation, 204
 - trigger action, 206
- Amplifiers, a-c power, 115
 - classes of operation, 116
 - equivalent circuit, 118
 - inductive loads with, 119
 - maximum power transfer, 119
- Amplifiers, a-c power, plate efficiency, 116
 - power transistors, 120–123
 - push-pull operation, 116
 - transformer coupling, 117
 - vacuum-tube type, 124–126
- d-c, 61–111
 - cathode-temperature variation, 65
 - chopper-stabilized, 97–103
 - definition, 61
 - design considerations, cathode-follower circuits, 108, 109
 - difference amplifier, dynamic load line, 110
 - static load line, 110
 - triode amplifiers, 103–108
 - with neon-tube-type coupling network, 106–108
 - with resistor coupling network, 103–106
 - drift, 61–80, 94, 95
 - ideal coupling network, 84
 - interstage coupling network, 80–84
 - linear analysis, 64
 - noise, 61
 - power-type, 124–126
 - with RC networks, 52–54
 - transistor, 84–97
- (*See also* specific types of amplifiers)
- Analogue, mechanical to electrical, 305–307
- Armature reaction, in Amplidyne, 187
 - in d-c motors, 213
 - definition, 187
 - effect of, in d-c generator, 180, 187
 - in d-c tachometer, 180
- Artificial horizon, 351, 353
- Askania nozzle valve, 420
- Aspirator, pneumatic, 475
- Asymptotic diagram, 14, 20
- Autosyn, 221
 - (*See also* Synchros)
- Bail ring, 347
- Ball, bouncing, 142

- Ball pump, 376
- Ball-screw actuator, 332, 333
- Base, transistor, 85
- Bellows, pneumatic, 455–457, 462, 476–480
 - common configurations, 480
 - differential-pressure type, 477
 - life expectancy, 479
 - material, 477
 - metals used, 478
 - summary of properties, 480
- Bernoulli force, 401, 470
- Bias battery, 81
- Bourdon tube, 472, 474, 476, 477
 - characteristics, 477
- Break point, 14
- Bridge, 21
- Bridged-T network, 43–46
 - with inductance, 53
 - as ripple filter, 254
 - use of, in carrier servo, 48
 - in mechanical networks, 308–312
- Brush-contact resistance, 206
- Brush-shift effect in Amplidyne, 201–203

- Cadillac valve, 427–429
- Capacitance, 20
 - effect on two-terminal-network function, 26
 - pneumatic, 445–447
- Capacitors, 89
 - active zone, 26, 27, 58, 59
 - air dielectric variable, 8
 - electrolytic, dry, 9
 - nonpolarized, 9
 - polarized, 9
 - mica, 9
 - oil-filled, 9
 - paper, 8
 - for silicon dielectric, 9
 - strain gauge, 340
 - tubular, 8
- Carrier-frequency servo, 48
 - effect of change in carrier frequency, 50–52
- Cascading of L sections, 24, 25, 42, 43
- Cathode bias, 70
- Cathode-coupled inverter, 79, 80
- Cathode follower, 25, 66–70
 - design, 108–109
 - drift, 67
 - due to component variation, 68
 - equivalent circuit, 69, 70
 - gain, 67
 - impedance, input, 67, 68
 - output, 67
- Cathode follower, phase inverter, 77, 78
 - with plate-load resistor, 69
- Cathode temperature variation (*see* Heater-voltage effect)
- Cavitation, hydraulic, 393
- Characteristic impedance of hydraulic line, 380
- Chézy flow formula, 377, 378
- Child's law, 415
- Chopper, 98, 99, 265, 274
 - electromechanical type, 98–99, 265–266, 274
 - short-circuiting type, 99
 - use in d-c amplifier, 64, 97–103
- Chopper frequency, 100, 101
- Cogging, 278
- Collector, transistor, 85
- Commutating poles, 187
- Complementary operation in transistor amplifiers, 126
- Complexity, system, 55, 56
- Compressibility, 435
 - as equivalent spring, 441
 - of hydraulic fluid, 368–369
- Compressible flow in pipes, 439–441
- Conduction period in thyratrons, 134
- Contact ratio of gears, 317
- Contraction coefficient of hydraulic line, 392
- Control transformer, synchro, 222–244
- Critical resistance, 193, 194
- Cylindrical rotors, 222

- Dashpot, 305
- D-c amplifier (*see* Amplifiers)
- D-c generator (*see* Generator)
- D-c motor (*see* Motors)
- D-c rate generator, 180, 181
- D-c tachometer, 180–181
- Deadband, 206, 388
 - in d-c motors, 213
 - in relays, 151, 152
- Decibels, 16
- Deltamax, 159
- Demodulator, 249–270
 - cathode-follower type, 261
 - clamping action, 267
 - clamping type, 266–268
 - maximum frequency, 268
 - peaking circuits, 268
 - describing functions, 152–154
 - diode, 261–265
 - electromechanical, 265–266
 - full-wave type, 254
 - diode type, 265
 - maximum frequency, 258

- Demodulator, full-wave type, output
 component, ripple, 257
 useful, 256
 superiority, 257
fundamental concepts, 249
half-wave type, 250–254
 analysis, 251
 diode type, 264
 harmonics, 252
 maximum frequency, 253
 output component, ripple, 252
 useful, 252
 phase shift, 252
 use of band-rejection filters, 254
inaccuracies, effect, 262
maximum frequency limit, 253, 258, 268
ring, 261
sampling action, 267
single-ended triode circuit, 259
triode type, 249–261
unbalance, 262
(*See also* Discriminator)
- Difference amplifier, 72
 design, 109–111
 effect of heater-voltage variation, 73
 equivalent form for, 73
 transistor type, 95
 vacuum-tube type, 72
- Differential network, 48
- Differential synchros, 241–243
- Differentiating gear trains, 334
 bevel-gear differential, 334
 spur-gear differential, 336
- Diode, 151
 free-wheeling, 135
- Discharge coefficient of hydraulic flow, 393, 419
- Discriminator, phase-sensitive, 224, 226, 231, 249–271
(*See also* Demodulator)
- Distribution factor, 229
- Dither, 390
- Drag-cup rotor, 278–280, 298
- Drift, of cathode follower, 67–69
 causes, 62
 in difference amplifier, 72–74
 effect of negative feedback, 62, 63
 in gyroscopes, 351–353
 in Miller circuit, 74–77
 in modulators, 273–274
 in single-stage triode amplifier, 71
 in transistor amplifiers, 94, 95
- Driving-point impedance, 19, 20
 RC , exact synthesis, 37
(*See also* Two-terminal network)
- E pick-off, 245–247
- Eddy current, 184–186
 effect in d-c generator, 184–186
- Efficiency of worm gear, 319, 320, 333
- Electromechanical chopper, 98–99, 265, 266, 274
- Electromechanical networks, 312–314
- Electromechanical vibrator (*see* Chopper)
- Electronic components, popularity, 1
- Emitter, 85
- Equivalent circuit, a-c servomotor, 281–283
 cathode follower, 69, 70
 heater-voltage effect, 65
 push-pull amplifier, 118
 triode, 65, 71
- Error, random, 220, 221
 repeatable, 220, 221
- Extinction point in thyratrons, 131–136
- Faraday's law, 161, 180
- Feedback, in chopper-stabilized d-c amplifier, 101–103
 effect on drift, 62, 63
 in magnetic amplifier, 171–175
 in relay amplifier, 150–157
- Figure of merit, a-c servomotor, 291–292
 d-c generator, 190–192
 magnetic amplifiers, 167, 171, 172
- Fire hazard, 435
- Firing angle of thyratrons, control, 127–129
 effect of inductive load, 129–138
- Flapper valves, 418–420
- Fliegner's equation, 439
- Flow through valve orifices, 392–397
- Flow separation, hydraulic, 393
- Flyweight tachometer, 341–343
 effect of nonlinearities, 343
 equations of motion, 342
 nonsinusoidal oscillation, 343
 spring-restrained type, 342
- Four-terminal networks, 18, 20–22
- Fourier analysis, 228, 251, 256, 270
- Free-wheeling circuit, 135, 136
- Frequency response from transfer function, 13
- Friction effect on d-c motor, 209
- Gear pump, 374–376
- Gear trains, 320–328
 backlash, 328, 330
 composite error, 328
 design for minimum inertia, 320–327

- Gear trains, differential gears, 334
 - power, 323–327
 - ratio for load matching, 327
 - in two-speed synchro systems, 329
- Gears, 315–333
 - backlash, 328–332
 - ball-screw actuator, 332
 - bevel gear and pinion, 318
 - Beveloid, 331–332
 - composite error, 328–329
 - contact ratio, 317
 - helical, 318
 - herringbone, 318
 - pitch, 318
 - pressure angle, 317
 - punched gear disk, 327
 - rack and pinion, 318
 - spring-loaded, 330
 - spur gear and pinion, 315
 - standards for fine-pitch service, 317
 - substitute materials, 327
 - tapered-tooth involute, 331
 - worm-and-gear, 318
 - efficiency, 319
 - maximum, 320
 - self-locking, 319, 320
- Generator, d-c, 179–207
 - band width, 190
 - figure of merit, 190, 191
 - power gain, 190
 - response speed, 190
- Goldberg amplifier, 102, 103
- Grooves in hydraulic control-valve pistons, 405
- Gyrocompass, 355–359
- Gyroscope, 343–361
 - applications, 353
 - free gyro, 353
 - gyroscopic velocimeter, 353, 354
 - to inertial navigation, 355–361
 - control of stable platform, 357
 - east-west errors, 358
 - effect of initial errors, 360
 - equations of motion of stable platform, 360
 - HIG gyro, 355–359
 - north-south errors, 358
 - simplified system block diagram, 359
 - transfer function, 349, 350, 355, 356
 - use of accelerometer with, 357
 - rate gyro, 354
 - restrained gyro, 353–355
 - approximate block-diagram representation, 348
 - basic torque equation, 344
- Gyroscope, cross-coupling transfer function, 350
 - definition of spin axis, 344
 - direct-transfer function, 350
 - equation of motion, 347
 - introduction, 343
 - linearized equations, 349
 - nutation, 349, 350
 - frequency, 349
 - practical gyros, 351
 - aircraft, 351
 - applications, 351
 - bearing friction, 351
 - marine gyrocompasses, 352
 - minimization of drift, 351
 - rate gyros, 352
 - precession in, 343–347
 - in conical surface, 348
 - Schuler-tuned, 352, 360
- h parameters, 88–94
- Hagen-Poiseuille law, 378, 379, 441
- Heater voltage, in difference amplifier, 73
 - effect on drift, 62
- Heater-voltage effect, equivalent circuit, 65
 - Miller circuit, 74–77
- Helical gear, 316, 318
- Herringbone gear, 318
- High-pass filter, 17
- Holes, 85
- Hooke joint, 337–339
- Hydraulic capacitance, 381
- Hydraulic control systems, 363–431
 - actuator, 386
 - unequal-area type, 387
 - analogy to Child's law, 414
 - axial reaction force, 397
 - transient, 399, 400
 - basic types, 363
 - bridge analogue for underlapped valve, 413
 - bulk modulus, 369
 - cavitation, 393
 - choice of operating pressure for, 431
 - compressibility coefficient, 369
 - compressibility flow, 368, 369
 - constant-flow system, 364
 - constant-pressure system, 364
 - control valves, axial reaction force, 397–402
 - flapper, 418
 - graphical analysis, 405–412
 - lap, 388
 - maximum power output, 407, 408
 - nozzle type, 420

- Hydraulic control systems, control
 valves, pressure-flow relations,
 392-397
 radial forces, 402-405
 slide, 420, 422
 small-signal linear analysis, 412-418
 spool type, 387-418
 two-stage, 423-429
definition, 343
equation of flow in underlapped region,
 396
flapper valve, 418
 balanced, 420
hydraulic accumulator, 430
 bellows type, 430
 hydropneumatic type, 430
hydraulic flow separation, 393
leakage flow, 368
linearized equivalent, for flow-control
 valve, 416
 stability, 417
linearized small-signal analysis, 412
lock, 402-405
maximum load power, 408
motor, 366-376
 flow, 368
pressure-regulating devices, 429-430
pump-controlled, 364, 365-367
 analysis 367-370
 transfer function, 370
pump flow, 368
pumps, 364-376
 ball, 276
 comparison of types, 370-376
 gear, 374-376
 piston, 365-372
 vane, 372-374
 variable-stroke, 366-372
reaction force, 397-405
 radial, 402
relief protection, 366
replenishing system, 366
slide valve, 420-422
stroke amplifier, 386
transmission, 367-370
transmission lines, 376
 absolute viscosity, 379
 characteristic impedance, 380
 Chézy flow formula, 377
 coefficient, of energy loss, 378
 of friction, 378
 kinematic viscosity, 379
 Reynolds number, 377
 turbulent flow, 377
 viscous or laminar flow, 376
two-stage valves, 423
 Cadillac, 427-429
- Hydraulic control systems, two-stage
 valves, feedback to remove inte-
 gration, 423
 feedback links, 424
 Moog, 425-427
 Siemen's, 424
valve controlled, 364, 365, 386-434
 Bernoulli force, 390, 401
 damping, 401
 flow-pressure relations, 392
 force-compensated, 402
 graphical analysis, 406
 introduction to, 386
 pulsed operation, 389
 advantages, 389-390
 disadvantages, 390
 spool-type valves, 387
 four-way, 388
 two-way, 387
 zero-lap, 388
 valve lock, 390, 402
 Venturi tube, 394
Hysteresis, magnetic, 181, 182, 184
 effect on phase lag, 184
 in relays, 152
Hysteresis curve in Amplidynes, 206
- Impedance, 13, 24, 25
 characteristic, of hydraulic lines, 380-
 382
Impedance analogue, 306
Impedance level in networks, 30, 34, 36,
 41
Impedance ratio in cascading of L sec-
 tions, 24, 42, 43
Inductance, of armature of d-c generator,
 187
 d-c motor armature effect in thyatron
 amplifiers, 142-144
 effect, on realizable transfer function of
 networks, 21, 22, 53
 in thyatron amplifier, 129-138, 142-
 144
 effective, in magnetic amplifiers, 166
Inductor, 1, 10, 11, 131
 air-cored, 10
 iron-cored, characteristics, 10
 never use, 11
 quality factor, 10
Inertial navigation, 351, 355-361
Infant mortality, 56
Input impedance of networks, 24, 25
 cathode-follower, 67, 68
 transducers, 221
 transistor-amplifier, 91-93
 triode-amplifier, 72

- Integral controller, 17
- Interstage coupling networks for d-c amplifiers, 80–84
- Kirchhoff's law, 183, 214
- Kirchhoff's mesh equation, 201
- L sections, 22–43
 - approximate synthesis procedure, 26–33
 - exact synthesis, general method, 34–43
 - simple transfer function, 33, 34
 - number required in ladder network, 25, 26
- Ladder networks, 24
 - approximate synthesis, 25–26
- Lag-lead network, 17
- Lag network, 17
- Lattice network, 21
- Lead network, 17, 48
- Lenz's law, 185
- Level changer, 83
- Linearity, best, 3
 - independent, 3
 - normal, 3
 - of variable resistors, 3
 - zero-based, 3
- Load line in hydraulic valves, 406
- Logarithmic representation of transfer function, 14
- Low-pass filter, 17
- Mach number, 440
- Magnesyn, 248
- Magnetic amplifier, 159–177
 - with feedback control, 171–175
 - gate winding, 159
 - hysteresis loop, 159
 - normally excited, 161
 - parallel-connected reactors, 167–171
 - figure of merit, 171
 - gain, a-c power, 169
 - current, 169
 - d-c power, 169
 - voltage, 169
 - time constant, 169–170
 - with positive feedback, 171
 - figure of merit, 172
 - gain, current, 171
 - d-c power, 172
 - voltage, 171
 - time constant, for falling transients, 172
 - for rising transients, 172
- Magnetic amplifier, Ramey, 175–177
 - shared-time principle, 177
 - self-saturated, 173–175
 - gain, current, 174
 - d-c power, 175
 - voltage, 174
 - time constant, for falling transients, 175
 - for rising transients, 175
 - series-connected type, 159–167
 - figure of merit, 167
 - gain, a-c power, 165
 - current, 164
 - d-c power, 164
 - voltage, 164
 - time constant, 165
 - use of special alloy, 159
- Magnetization curve, 182, 193
- Mass, 305, 308
- Mechanical filters, 305–315
- Metadyne, 199
- Microsyn, 247, 248
- Miller circuit, 74–77
 - effect of change in positive supply voltage, 77
 - equivalent circuit for, 75
 - practical form, 76
- Mobility, analogue, 305–315
- Modular packing, 57
- Moog valve, 425–427
- Motors, d-c, 207–218
 - analysis inaccuracies, 213
 - armature reaction, 213
 - brush-contact resistance, 213
 - deadband, 213
 - electrical damping, 209, 215, 217
 - equivalent circuit, 212
 - field-controlled type, 216–218
 - gain constant, 218
 - mechanical output impedance, 209
 - quasi-linear transfer function, 211
 - relation, between constants, 208
 - between transfer function and speed-torque curves, 210
 - separate excitation, 207–213
 - speed regulation, 142
 - split-field series type, 213–218
 - thyatron-controlled, 139–149
 - time constant, 210
- pneumatic, 474–476
 - gear type, 475
 - piston type, 476
 - diesel action in, 476
 - vane type, 475
- Mutual inductance in d-c generators, 187–190

- Negative feedback effect on drift, 62, 63
Neon tubes for interstage coupling in d-c amplifiers, 81, 106–108
Node method, 11
Nonbleed relay, 474
Nonbleed valve, 474
Normal excitation, 161
Normal law, 220
Null networks, 43–52
Nutation of gyroscope, 349, 350
Nylon used for gears, 327
Nyquist diagram, 153–154
- Operating pressure of pneumatic systems, 467
Orifice coefficient, 439
Oscillations in thyatron circuits, 150
Output impedance, Amplidyne, 202–205
 cathode follower, 67
 d-c generator, 186–190
 mechanical, in d-c motor, 209
 networks, 24, 25
 transistor amplifier, 91–93
 triode amplifier, 71
Overlapped valves, 388
- Peaking circuits, 268
Phase inverter, 77–80, 124, 125
 balanced, 78
 cathode-coupled type, 79
 cathode-follower type, 77
 paraphase type, 78
Phase-sensitive detector (*see* Discriminator)
Pickup device, 220, 221
Pinion, 315–317, 322
Piston pumps, 365–372
Pitch factor, 229
Pneumatic actuator, 476–480
 differential area, 476
 linear, 476
 spring-loaded, 476
 (*See also* Bellows; Bourdon tube)
Pneumatic amplifier, 455
Pneumatic aspirator, 475
Pneumatic bellows, 276–280, 455–457, 462
Pneumatic capacitance, 444–447
Pneumatic components, 461–480
 interchangeable with hydraulic, 461
 introduction to, 461
 power supplies, 467–468
 advantages of multistage compression, 468
 weight advantage over hydraulic, 465
Pneumatic components, weight comparison with electric and hydraulic, 463
Pneumatic compressor, 467
 advantages of multistage compression, 468
Pneumatic-control functions, 443
Pneumatic-control valve, 470–474
 effect of viscosity, 470
 two-stage flapper-control pressure valve, 473, 474
 underlap characteristic, 472
Pneumatic-electric analogues, 443–447
Pneumatic equalizer, 463
Pneumatic-equalizer networks, 447–454, 463
Pneumatic flapper-type control valves, 453, 454, 456, 470, 473, 474
Pneumatic flow, 437–439
 adiabatic, 437
 compressible, in pipes, 439
 Mach number, 440
 pipe friction factor, 440
 relation for, 439
 Reynolds number, 440
 critical-pressure ratio, 439
 general energy equation, 437
 through orifice, 438
Pneumatic jet-engine control servo, 461–463
Pneumatic motors (*see* Motors)
Pneumatic orifices, 438, 439
Pneumatic relay, 455
Pneumatic resistance, 444–447
Pneumatic spool-type control valves, 454–456, 470–474
Pneumatic systems, 435–480
 advantage over hydraulic system, 435
 choice of operating pressure for, 467
 disadvantages as compared to hydraulic, 436
 explosion hazard, 436, 465, 467
 introduction to, 435
 response speeds, 437
Pneumatic transmission lag, 441
 first-order approximation for, 442
Pneumatic valve, equivalent circuit, 456
Polarized relay, 216
Pole face windings, 187
Poles, effect of successive, 32, 33, 36
 of network function, 19–25
 networks with complex, 52–54
 restrictions on, 21–25
Positive feedback in magnetic amplifiers, 171–173
Potentiometer (*see* Resistors, variable)
Potentiometer slide wire, 5

- Power amplifiers, d-c, 123-126
 - single-ended push-pull, 124, 125
 - transistor, 125
- Precession, 343-347
- Precision, 220, 221
- Pressure-regulating devices, 429, 430
- Propagation constant of hydraulic line, 380
- Proportional plus rate network, 17
- Proportional plus rate plus integral controller, 17
- Proportional plus reset network, 17
- Pulse-length modulation, 389-392
- Pulsed operation of hydraulic valves, 389-392
- Push-pull amplifier, ideal, 126

- Q in inductor 10, 11
- Quadrature in strain gauges, 341
- Quadrature circuit in Amplidyne, 199-206
- Quadrature voltage in synchros, 225-226

- Radio Manufacturers Association, 2
- Ramey magnetic amplifier, 175-177
- Rate generator, 180-181
- Rate gyro, 354, 355
- RC networks, 11-54
 - amplifiers, 52-54
 - use of, 52
 - bridged-T, 43
 - five simple, 17
 - four-terminal, 20
 - general properties, 18
 - L section, 22
 - approximate synthesis, 26
 - cascading, 24
 - exact-synthesis method, 33-34
 - ladder structure, 22
 - lattice structure, 22
 - synthesis, 17-43
 - introduction to, 17
 - table of common types, 17
 - twin-T, 43
 - two-terminal, properties, 19
- RC rate network, 11, 17
- Reference voltage, 224, 249-250, 276
- Regulex, 180, 197-198, 207
- Rejection amplifier, 53
- Rejection network (*see* Twin-T circuits)
- Relay amplifiers, 150-159
 - chatter rate, 156
 - deadband, 151
 - describing function, 152
- Relay amplifiers, frequency response, 156-158
 - hysteresis, 152
 - output pulse-repetition rate, 156
 - polarized relay, 151, 213
 - stability, 152-154
 - static characteristic, 154-156
- Reliability, 54-57
- Relief valve, 366, 367
- Replenishing system, 366, 367
- Reset network, 17
- Residual voltages in synchros, 225, 226
- Resistance, 20
 - as interstage coupling in d-c amplifiers, 83, 104-106
 - pneumatic, 444-447
 - strain gauge, 340-341
 - temperature coefficient, 7, 8
- Resistance-capacitance networks (*see* RC networks)
- Resistance padding, 6, 7
- Resistors, 1-8
 - carbon, 1
 - color code, 1
 - high-resistance wire, 1
 - linearity, 3
 - precision variable, 3
 - mechanical characteristic, 5
 - nonlinear, 6, 7
 - resistance padding, 6
 - voltage padding, 6
 - slide-wire, 5
 - standard power ratings, 1
 - standard tolerances, 1
 - standard values, 2
 - variable, 2-7
 - gauging, 6
 - linearity, 3, 4
 - moment of inertia, 5
 - resolution, 4
 - starting torque, 5
- Resolution, 4
- Resonance, 15
- Reynolds number, 377, 378, 440
- Ripple, 180, 252-262
 - commutator, 181, 300
 - in discriminator output, 252-262
- Ripple filters, 253-259
- Root form of transfer function, 12
- Rosenberg generator, 199
- Rotors for synchros, 222
- Rototrol, 180, 192-197, 207
 - analysis, 192
 - critical resistance, 193
 - critically tuned, 194
 - definition, 192
 - effect of eddy currents in, 195

- Rototrol, with multiple-control windings, 197
nonminimum phase transfer, 196
pilot generator, 192
transfer function, 195
- Salient pole rotor in synchros, 222, 227, 228
- Sampled-data analysis, 270
- Saturation curve, 181
- Schuler tuning of gyroscopes, 352, 360
- Self-saturation in magnetic amplifiers, 173-175
- Selsyn, 221
(*See also* Synchros)
- Separation principle of poles and zeros, 20
- Servomotors, a-c, 49, 276-303
advantages, 277
approximate transfer function, 288
characteristics with finite control impedance, 292
cogging, 278
construction features, 277
control by thyratrons, 149
cross-field theory, 299
double-revolving-field theory, 299
effect of unbalanced stator winding, 294
efficiency, 277
equivalent circuit, 282
figures of merit, 29
impedance in control circuit, 292-294
as phase discriminator, 277, 286
power range, 277
resonating capacitor, 277
rotor types, 278
shaded-pole induction motor, 300-303
single-phasing, 290, 296-298
slip, 284
slot lock, 278
speed-torque curves, 288-295
starting torque, 286
symmetrical components, 284
synchronous speed, 281
theory of operation, 279
use, 276-277
as tachometer, 298
- Shaded-pole a-c motor, 300-303
- Shared-time principle, 177
- Single phasing of a-c servomotors, 290, 296-298
- Skin effect in a-c servomotor, 286-287
- Sleeve valve, 424
- Slewing, 237
- Slide valve, 420-422
- Slip, 281
- Slot lock, 278
- Smith chart use with hydraulic transmission, 380
- Solid iron rotor, 287
- Space harmonics, in a-c servomotors, 277
in synchro air-gap flux, 229
- Sparkling, 187, 199, 200
- Spring, 305
- Spur gear, 315-317
- Square-wave modulator, advantages, 258
transistor circuit, 272
- Square waves, 251-274
- Squirrel-cage rotor, 278
- Stability of relay amplifiers, 152-154
- Stable platform, 357-359
- Stator of synchros, 222
- Stick-off voltage, 240, 241
- Strain gauge, 339-341
- Supermalloy, 159
- Suppressed-carrier amplitude modulation, 224, 249, 251, 271
- Symmetrical components in analysis of a-c servomotor, 284-285
- Synchro capacitors, 243
- Synchros, 49, 115, 220-244
classification, 222
construction, 222
control transformer, 222-244
differential units, 241-243
elementary operation, 221
generator, 222-244
harmonic voltage component, 226
N-speed systems, 235
one-speed systems, 235
output-signal form, 49, 223, 224
quadrature error, 225-226
repeater, 222, 243-244
static errors, 225, 227-229
suppressed-carrier amplitude modulation, 224
synchro capacitor, 243
synchronous velocity, 234
torque gradient, 244
two-speed systems, 235-241
diode-clipper switching circuit, 237
false point synchronization, 238
neon-tube switching circuit, 236
oscillation, 238
slewing characteristics, 237
stick-off voltage, 240
switching circuits, 236-241
velocity error, 227, 230-235
vibration damper, 243
- Synchronization, 236
- Sziklai amplifier, 126

- Tachometer, a-c type, 298–300
armature reaction, 180
brush bounce, 180
d-c type, 180–181
flyweight, 341–343
ripple component, 180
- Teleon, 248
- Temperature coefficient of resistance, 7, 8
- Temperature effect, on thynatron conduction, 127
on transistors, 84, 87, 94
- Thermal runaway, 94
- Thermocouple, 97
- Thévenin circuit, 182, 186, 207
of Amplidyne, 202–205
of d-c motor, 207
equivalent mechanical, 209
- Three-terminal networks, 18, 20
- Thyratron, conduction, 126
control with, of a-c motors, 149
of firing angle, 127–129
of split-field series motor, 149
grid potential, 126, 127
operation, 126
positive ion sheath, 126
tube drop, 127, 132–135
- Thyratron amplifier, 126–150
continuous load current, 136–137
with d-c motor load, 139–149
bidirectional control, 145–148
extinction angle, 131–143
firing angle, 128, 133, 134, 135, 140
free-wheeling circuit, 135
full-wave, 136, 137
half-cycle response time, 138
with inductive loads, 129–138
one-cycle response time, 138
reactance tube circuit, 129
transfer function, 138
tube drop, 140
- Time constant, Amplidyne, 205
d-c motor-controlled thyratron, 148
magnetic amplifiers with feedback, 172
motor inertia, 210
parallel-connected magnetic amplifier, 169–171
self-saturated magnetic amplifier, 175
series magnetic amplifier, 165–167
- Time-constant form of transfer function, 12
- Torque, a-c servomotor, 49, 283–298
- Torque gradient in synchro repeater, 244
- Torque to inertia ratio of a-c servomotors, 279, 291
- Torque motor, hydraulic, 390, 429
- Transducers, 220
accuracy, 221
errors, 220
precision, 221
variable reluctance, 245
- Transfer function, a-c servomotor, 288–291
all-pass type, 54
Amplidyne, 203–205
with complex zero, 44–46
d-c generator, 189, 190
definition, 11
field-controlled d-c motor, 217, 218
frequency characteristic, 23
gyroscope, 349, 350, 355, 356
hydraulic transmission, 370
with imaginary zeros, 46–52
nonminimum phase, 53
pneumatic bellows, 455–457
RC networks with amplifiers, 52–54
realizable, with four-terminal *RC* network, 20
root form, 12
Rototrol, 194–196
separately excited d-c motor, 209–212
split-field series motor, 215
steady-state frequency response from, 13
synchro, 233–235
thyratron amplifier, 138
quasi-linear, 138, 142
time-constant form, 12
- Transient hydraulic-reaction force, 399–401
- Transistor amplifiers, complementary
operation, 126
drift, 94, 95
power-type, 120–123
- Transistor circuits, difference amplifier, 95
grounded base, 86, 91–93, 120
grounded collector, 91–93, 121
grounded emitter, 87–93, 121
thermal runaway, 94, 120
- Transistors, 84–97, 120–123, 272–273
action defined, 85
analysis of simple circuits using, 85–90
base, 85
collector, 85
complementary amplifier, 125, 126
effect of temperature on, 84, 87, 94
emitter, 85
equivalent circuits, 86
h parameters, 88–94
modified, 93
holes in, 85
junction, 85

- Transistors, modulators, 272, 273
 multistage amplifiers, 95-97
 single-stage circuits, 90-94
 stability factor, 94
 temperature dependence of parameters
 in, 87
 temperature-sensitive elements for
 compensation, 95
 use of, in a-c power amplifiers, 120-123
 in d-c amplifier, 84
 in d-c power amplifier, 125
- Triode, as cathode resistor, 80
 characteristic curves, 64
 equivalent circuit, 65
- Triode circuits, 64-80
 single-stage amplifier, 70-72
 drift, 71
 gain, 71
 grounded-grid amplifier, 72
 output impedance, 71
- Twin-T circuits, use, with amplifiers in
 generation of complex pole-transfer functions, 52-54
 in carrier servo, 48
 in mechanical network, 308, 312
 as series equalizer in a-c servos, 48-52
- Two-terminal network, 18, 19
 effect of capacitors, 26
 exact synthesis, 37, 38
 properties, 19, 20
- Two-terminal *RC* network, approximate
 synthesis method, 27, 28
- Umbrella rotor, 222
- Underlapped valves, 388, 395-397
- Unequal-area hydraulic actuator, 387
- Universal joint, 337-339
 velocity error, 339
- Vacuum tubes, reliability, 55, 57
- Valve-controlled hydraulic systems (*see*
 Hydraulic control systems)
- Valve lap, 388
- Valves, hydraulic (*see* Hydraulic control
 systems, control valves)
- Vane pumps, 372-374
- Variable-reluctance transducers, 245-248
- Variable-reluctance transformer, 247
- Velocity coefficient of hydraulic flow, 392
- Velocity effect on synchro gain, 234, 235
- Velocity error, 227, 230-235
- Velocity propagation in hydraulic lines,
 380
- Vena contracta, 298
- Venturi tube, 394
- Vibration damper in synchro repeaters,
 243
- Virtual cathode, 65
- Viscosity, absolute, 379
 kinematic, 379
- Voltage-divider principle of obtaining
 transfer function, 12, 13, 22
- Voltage padding, 6, 7
- Ward Leonard system, 179, 207, 216
- Wear-out region, 56
- Wobble plate, 365
- Worm gear, 318-322
- Zener diode, 82-83, 97
- Zero-lapped hydraulic valves, 388, 406-412
- Zeros, complex, network with, 44-52
 of network function, 19, 20, 22-25
 purely imaginary, network with, 44, 46-49
 restrictions on, 22-25

